UNIVERSITY OF CALIFORNIA
RIVERSIDE

Understanding How Lexical and Multisensory Contexts Support Speech Perception

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Psychology

by

Joshua Jeremiah Dorsi

September 2019

Dissertation Committee:
Dr. Lawrence D. Rosenblum, Chairperson
Dr. Aaron Seitz
Dr. Curt Burgess

The Dissertation of Josh Jeremiah Dorsi is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

Acknowledgements

I give my most sincere thanks to my Committee Chair, Lawrence D. Rosenblum. Under Larry's mentorship I learned to pursue the questions that most interested me even when they were challenging and complex. Larry was always an invaluable source of knowledge, of genuine enthusiasm for research, and patient feedback. These are some of the qualities that made Larry a superb advisor and that I hope to develop within myself.

I also give deep thanks to Curt Burgess and Aaron Seitz who each provided invaluable insight and perspective in the development of this dissertation. Beyond their roles as committee members, both Curt and Aaron were appreciated sources of guidance throughout my graduate education.

I also thank Arthur Samuel and Rachel Ostrand whose work motivated parts of this dissertation and who were fantastic collaborators. Both Arty and Rachel taught me a great deal, and I hope to continue working with them in the future.

My success was also supported by many of the other faculty members of the UCR Department of Psychology. In particular Glenn Stanley was a great source of advice that I am very fortunate to have had.

I am also thankful for the support of my colleagues at the UCR Graduate Writing Center. I am especially thankful for Hillary Jenks and Christina Trujillo who were compassionate and supportive during a time when I needed it.

Finally, I extend my sincere thanks to faculty from the SUNY New Paltz Psychology Department where my graduate career began. In particular I wish to thank Navin Viswanathan who has continued to be a source of guidance.

Dedication

This dissertation and much of my graduate career would not have been possible without the love and support of my wife Nichole Mueller. For more than a decade you have been my best friend. You inspire me to do more, to work harder, and to be better. I emulate your wit, your compassion, and your grace. Together we have gone on many adventures and shared many perfect moments. We have also endured incredible hardships, perhaps none as trying as the heartbreak and turmoil of these last few months, but through it all we have stuck together. I am grateful for every minute we have been together and I cannot wait to meet the next chapter of our life together.

ABSTRACT OF DISSERTATION

Understanding How Lexical and Multisensory Contexts Support Speech Perception

by

Joshua Jeremiah Dorsi

Doctor of Philosophy, Graduate Program in Psychology
University of California, Riverside, September 2019
Dr. Lawrence D. Rosenblum, Chairperson

The perception of speech is supported by both multisensory (e.g. Sumby & Pollack, 1954) and lexical information (e.g. Miller, Heise, & Lichten, 1951). How the mechanism for speech perception processes these two sources of information is an open theoretical question. Evidence indicates that multisensory information is integrated early in speech processing (e.g. Musacchia et al., 2006) and some theories assume that integration precedes lexical processing (e.g. Fowler, 2004; see also Rosenblum et al., 2016). Accordingly, these theories assume that lexical processing is performed on the integrated multisensory information. In contrast, some have recently proposed that lexical processing is performed on unintegrated unisensory information (e.g. Ostrand et al., 2016; Samuel & Lieblich 2014). This dissertation provides a careful investigation into these claims to address the potential interactions between lexical processing and

multisensory integration. Chapter 1 investigates if semantic processing of McGurk stimuli is consistent with the unperceived (and putatively unintegrated) auditory information (i.e. Ostrand et al., 2016) or the perceived (audio-visually integrated) information. Chapter 2 investigates if selective adaptation, a perceptual phenomenon known to be sensitive to low-level sensory information (e.g. Samuel & Newport, 1979) but also to lexically supported illusory percepts (e.g. Samuel, 1997), is sensitive to multisensory illusions. Finally, Chapter 3 investigates if lexical information influences the integration of auditory and visual speech information. The results of this dissertation indicate that lexical processing is sensitive to integrated multisensory information. However, this dissertation found no indication that lexical information influenced the multisensory integration process.

Table of contents

List of Figures

List of Tables

Introduction

Speech consists of *words* spoken by a talker that can be both *seen and heard*; that is, speech has both *lexical* and *multisensory* information. While the supportive effects of lexical and multisensory information on speech perception were first reported more than half a century ago (Miller, Heise, & Lichten, 1951; Sumby & Pollack, 1954), research has generally studied their effects in isolation from one other. In the multisensory literature, there is evidence that cross-sensory information is completely integrated early in perceptual processing (see Rosenblum, Dorsi, & Dias, 2016 for a review). However, within the few studies that have examined both lexical and multisensory contexts, some recent research indicates that lexical information influences speech perception independent of multisensory integration (e.g. Samuel & Lieblich, 2014; Ostrand et al., 2016). Thus, while multisensory information is thought to be integrated early, lexical processing seems to sometimes operate on unisensory information.

In light of this paradox, this dissertation will investigate the interactive processing of lexical and multisensory information in a series of three projects, presented here in three separate chapters. Each chapter will address a question motivated by a different account that has been put forward for the processing of multisensory and lexical information. However, the scope of these chapters extends beyond these motivating accounts. Within each chapter, we discuss multiple theories relevant to the tested research question and test multiple competing predictions.

Speech is Multisensory

Some of the earliest examples of the multisensory nature of perception have come from speech. Perhaps the most well known of these examples is the McGurk effect, the finding that discrepant visual speech can alter how auditory speech is heard (McGurk & MacDonald, 1976). For example, auditory 'ba' + visual 'ga' is sometimes heard as the visual stimulus, 'ga', or a fusion of the auditory and visual stimuli, such as 'da' (MacDonald & McGurk, 1978). Even before this seminal study, research had established that congruent visual speech could facilitate auditory speech perception (Sumby & Pollack, 1954). In the decades since these classic studies, there has been accumulating evidence for just how fundamental multisensory information is to speech perception (See Rosenblum et al., 2016 for a review).

Early Multisensory Speech Processing

Much research has investigated the time course of multisensory integration (for a review, see Rosenblum et al., 2016). This research has demonstrated that multisensory information can influence speech perception as early as feature recovery (Brancazio, Miller, & Paré, 2003; Fowler, Brown, & Mann, 2000; Green & Kuhl, 1989; Green & Miller, 1985). Other research shows influences of visual speech in auditory brain areas, often occurring too soon following visual stimulus onset to have been produced by feedback from other brain areas (See Besle, Fort, Delpuech, & Giard, 2004; Besle et al., 2008). There is also evidence of visual speech modulating auditory processing as early as the brainstem (Musacchia, Sams, Nicol, & Kraus, 2006). Overall, it seems that cross-sensory processing occurs early in perception.

Multisensory Speech Processing Is Encapsulated

Not only does it seem that multisensory integration occurs early, but there is also evidence that this process is encapsulated from top-down factors. For example, research using the McGurk effect shows that multisensory integration occurs even when participants are made aware of the incongruent nature of the stimuli (Summerfield & McGrath, 1984). Similarly, directing attention to one stimulus modality also does not diminish the McGurk effect (Massaro, 1987). The McGurk effect persists when the auditory and visual stimuli originate from speakers of different genders (Green et al., 1991); and when the auditory and visual signals are desynchronized or come from very different locations (Munhall, Gribble, Sacco, & Ward, 1996; Jones & Munhall, 1997; Jones & Jarick, 2006).

Multisensory Speech Processing is Automatic

There is evidence of multisensory integration occurring across stimuli that are rarely experienced together; auditory and tactile speech stimuli can integrate to improve accuracy for speech in noise listening (Gick, Johannsdottir, Gibraiel, & Mühlbauer, 2008; Sato et al., 2010) or produce McGurk like effects (Gick & Derrick, 2009; Fowler & Dekle, 1991). There are also demonstrations of audio-visual speech integration when visual speech is presented outside the awareness of the listener (Rosenblum & Saldana, 1996; Munhall, ten Hove, Brammer, & Pare, 2009).

Linguistic Speech Processing Is Multisensory

In light of the above discussed literature, it may be unsurprising that linguistic processing of speech is also multisensory. Consistent with the assertion that the speech brain is blind to modality, visual speech seems to mirror the lexical organization of auditory speech, showing effects of lexical frequency and linguistic neighborhoods (Tye-Murray, Sommers, & Spehar, 2007; Strand & Sommers, 2011). Moreover, visual speech influences the identification of basic phonetic features (Brancazio et al., 2003; Fowler et al., 2000; Green & Kuhl, 1989; Green & Miller, 1985). Visual speech can also facilitate the processing of complicated speech (Arnold & Hill, 2001; Bernstein, Auer, & Takayanagi, 2004; Reisberg, McLean, & Goldfield, 1987). Finally, there is also evidence that linguistic processing can cross modalities. Two studies, Kim et al., (2004) and Fort et al., (2013) have shown evidence of visual-only speech interacting with the lexical processing of auditory speech. In short, there is evidence that linguistic processing is sensitive to multisensory speech information.

Models of Multisensory and Linguistic Processing

In light of the apparent ubiquity of multisensory information in speech perception, there have been a number of attempts to formulize the interaction of linguistic and multisensory processing in a single framework. This dissertation focuses on three such accounts.

Lexical Processing Begins Before Multisensory Integration is Complete

Ostrand et al., (2016) propose that lexical processing and multisensory integration occur in parallel, with the lexical information from the auditory component being processed before multisensory integration completes (Figure 0.1). This account was motivated by a finding that seems to suggest that the unperceived auditory component of McGurk stimuli drives semantic priming (Ostrand et al., 2016; but see Chapter 1). The key assumption of this account is the privileged access of auditory speech in linguistic processing. Under this account, when audio-visual speech enters the speech process, the auditory speech is immediately analyzed for linguistic meaning. In contrast, this account assumes that before visual speech can be linguistically analyzed it must be integrated with the auditory signal. Thus, under this account, auditory speech is always linguistically processed faster than visual speech. This assumption leads to interesting predictions in McGurk contexts; under this account a perceiver might illusorily "hear" the visual speech but linguistically process the unperceived auditory signal.

The Linguistic Process is Independent of Multisensory Perception

Samuel and Lieblich (2014) suggest that lexical and multisensory contexts influence speech perception through separate and independent processes (Samuel & Lieblich, 2014). Under this account, one perceptual pathway integrates multisensory information and determines the phenomenological experience of a speech stimulus. Concurrent with this processing, a separate pathway processes the linguistic information from a speech stimulus. These authors argue that lexical context has far reaching effects

on even the most fundamental linguistic and perceptual processes, while multisensory context is limited to superficial, non-linguistic, perceptual processes (See Figure 0.2).

This account is similar to the one offered by Ostrand et al., (2016): both accounts propose an early dissociation between the perception of speech and the linguistic processing of it. There are, however, key differences between these accounts that should be noted. First, the account of Ostrand et al., (2016) was put forward specifically to address a dissociation observed in semantic priming; this account does not require that this dissociation operate at pre-lexical levels of processing. Second, and more important, the Ostrand et al., (2016) model is a time sensitive account. The authors never claim that multisensory information cannot access linguistic processing, only that, at least with respect to semantic processes, acoustic information will be processed *before* multisensory information.

In contrast, Samuel and Lieblich (2014) argue that even the most basic (i.e. featural) levels of linguistic processing are independent of multisensory perception. In fact, these authors offer no mechanism by which multisensory perception might gain access to these linguistic processes. Samuel and Lieblich (2014) support their claim with the observed dissociation between lexical and multisensory influences on *selective adaptation*. Interestingly, lexical, but not multisensory, context influences selective adaptation (see Samuel & Lieblich, 2014). Samuel and Lieblich (2014) argue that this dissociation in selective adaptation reflects a distinction in the processing of speech, with multisensory information influencing the phenomenological experience of speech, but not the linguistic processing of that speech.

Multisensory Integration Does Not Process Lexical Information

Brancazio (2004) formulates a model in which multisensory speech identification involves two stages. In the first stage, cross-sensory inputs are combined. The second stage concerns the phonetic categorization of that integrated output (See Figure 0.3). Here, we should note that these points for lexical influences on perception are not necessarily mutually exclusive. Brancazio (2004) notes that linguistic processing could take place during either, or both, of these stages. In contrast, a prominent theory of perception, the amodal account, (Fowler, 2004; see also Rosenblum et al., 2016), predicts that lexical processing will be restricted to post integration stages of processing. Chapter 3 investigates the stages of perception and their sensitivity to lexical information. The results are interpreted with respect to the Brancazio (2004) framework as well as their implications for the amodal account.

In the following sections we will empirically test predictions formed by each of these three models and compare them to predictions formed by competing accounts. In the discussion section we will discuss the broader implications of the three chapters taken together.

References

Arnold, P., & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology*, *92*(2), 339–355. http://doi.org/10.1348/000712601162220

Bernstein, L. E., Auer, E. T., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication*, *44*(1–4 SPEC. ISS.), 5–18. http://doi.org/10.1016/j.specom.2004.10.011

Besle, J., Fischer, C., Lecaignard, F., Bidet-Caulet, A., Lecaignard, F., Bertrand, O., & Giard, M. (2008). Visual activation and audiovisual interactions in the auditory cortex during speech perception : Intracranial recordings in humans. *The Journal of Neuroscience*, *28*(52), 14301–14310. http://doi.org/10.1523/JNEUROSCI.2875-08.2008

Besle, J., Fort, A., Delpuech, C., & Giard, M. H. (2004). Bimodal speech: Early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience*, *20*(8), 2225–2234. http://doi.org/10.1111/j.1460-9568.2004.03670.x

Brancazio, L. (2004). Lexical influences in audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(3), 445–463. http://doi.org/10.1037/0096-1523.30.3.445

Brancazio, L., Miller, J. L., & Paré, M. A. (2003). Visual influences on the internal structure of phonetic categories. *Perception & Psychophysics*, *65*(4), 591–601. http://doi.org/10.3758/BF03194585

Fort, M., Kandel, S., Chipot, J., Savariaux, C., Granjon, L., & Spinelli, E. (2013). Seeing the initial articulatory gestures of a word triggers lexical access. *Language and Cognitive Processes*, *28*(8), 1–17. http://doi.org/10.1080/01690965.2012.701758

Fowler, C. A. (2004). Speech as a supramodal or amodal phenomenon. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *Handbook of Multisensory Processes* (pp. 189–201). Cambridge.

Fowler, C. A., Brown, J. M., & Mann, V. A. (2000). Contrast effects do not underlie effects of preceding liquids on stop-consonant identification by humans. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(3), 877–888. http://doi.org/10.1037//O096-1523.26.3.877

Fowler, C. A., & Dekle, D. J. (1991). Listening with eye and hand: Cross-modal contributions to speech perception. *Journal of Experimental Psychology. Human Perception and Performance*. http://doi.org/10.1037/0096-1523.17.3.816

Gick, B., & Derrick, D. (2009). Aero-tactile integration in speech perception. *Nature*, *462*(7272), 502–504. http://doi.org/10.1038/nature08572

Gick, B., Jóhannsdóttir, K. M., Gibraiel, D., & Mühlbauer, J. (2008). Tactile enhancement of auditory and visual speech perception in untrained perceivers. *The Journal of the Acoustical Society of America*, *123*(4), EL72-6. http://doi.org/10.1121/1.2884349

Green, K. P., & Kuhl, P. K. (1989). The role of visual information in the processing of place and manner features in speech perception. *Perception & Psychophysics*, *45*(1), 34–42. http://doi.org/10.3758/BF03208030

Green, K. P., Kuhl, P., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & Psychophysics*, *50*(6), 524–536. http://doi.org/10.3758/BF03207536

Green, K. P., & Miller, J. L. (1985). On the role of visual rate information in phonetic perception. *Perception & Psychophysics*, *38*(3), 269–276. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/4088819

Jones, J. A., & Jarick, M. (2006). Multisensory integration of speech signals: The relationship between space and time. *Experimental Brain Research*, *174*(3), 588–594. http://doi.org/10.1007/s00221-006-0634-0

Jones, J. A., & Munhall, K. G. (1997). Effects of separating auditory and visual sources on audiovisual integration of speech. *Canadian Acoustics*, *25*(4), 13–19.

Kim, J., Davis, C., & Krins, P. (2004). Amodal processing of visual speech as revealed by priming. *Cognition*, *93*(1). http://doi.org/10.1016/j.cognition.2003.11.003

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics*, *24*(3), 253–7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/704285

Massaro, D. (1987). *Speech Perception by ear and eye: A paradigm for psychological inquirey*. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.

Miller, G. A., Heise, G. A., & Lighten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *The Journal of Experimental Psychology*, *41*(5), 329–335.

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics*, *58*(3), 351–362. http://doi.org/10.3758/BF03206811

Munhall, K. G., ten Hove, M. W., Brammer, M., & Paré, M. (2009). Audiovisual integration of speech in a bistable illusion. *Current Biology : CB*, *19*(9), 735–9. http://doi.org/10.1016/j.cub.2009.03.019

Musacchia, G., Sams, M., Nicol, T., & Kraus, N. (2006). Seeing speech affects acoustic information processing in the human brainstem. *Experimental Brain Research*, *168*(1–2), 1–10. http://doi.org/10.1007/s00221-005-0071-5

Ostrand, R., Blumstein, S. E., Ferreira, V. S., & Morgan, J. L. (2016). What you see isn't always what you get: Auditory word signals trump consciously perceived words in lexical access. *Cognition*, *151*, 96–107. http://doi.org/10.1016/j.cognition.2016.02.019

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A speechreading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97–113). London, England: Erlbaum.

Rosenblum, L. D., Dias, J. W., & Dorsi, J. (2016). The supramodal brain: Implications for auditory perception. *Journal of Cognitive Psychology*, *5911*, 1–23. http://doi.org/10.1080/20445911.2016.1181691

Rosenblum, L. D., Dorsi, J., & Dias, J. W. (2016). The impact and status of Carol Fowler's Supramodal Theory of Multisensory Speech Perception. *Ecological Psychology*, *28*(4), 262–294. http://doi.org/10.1080/10407413.2016.1230373

Rosenblum, L. D., Johnson, J. A., & Saldana, H. M. (1996). Point-light facial displays enhance comprehension of speech in noise. *Journal of Speech & Hearing Research*, *39*(6), 1159. http://doi.org/10.1044/jshr.3906.1159

Samuel, A. G., & Lieblich, J. (2014). Visual speech acts differently than lexical context in supporting speech perception. *Journal of Experimental Psychology. Human Perception and Performance*, *40*(4), 1479–90. http://doi.org/10.1037/a0036656

Sato, M., Cavé, C., Ménard, L., & Brasseur, A. (2010). Auditory-tactile speech perception in congenitally blind and sighted adults. *Neuropsychologia*, *48*(12), 3683–3686. http://doi.org/10.1016/j.neuropsychologia.2010.08.017

Strand, J. F., & Sommers, M. S. (2011). Sizing up the competition: Quantifying the influence of the mental lexicon on auditory and visual spoken word recognition. *The Journal of the Acoustical Society of America*, *130*(3), 1663–1672.

http://doi.org/10.1121/1.3613930

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*(2), 212–215.

Summerfield, & McGrath. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. *Quarterly Journal of Psychology*, *36A*, 51–74.

Tye-Murray, N., Sommers, M., & Spehar, B. (2007). Auditory and visual lexical neighbourhoods in audiovisual speech perception. *Trends in Amplification*, *11*(4), 233–241. Retrieved from https://docs.google.com/file/d/0B3N-C9xfG_CPbWNEQ3pKREhWU2M/edit

Figure 0.1

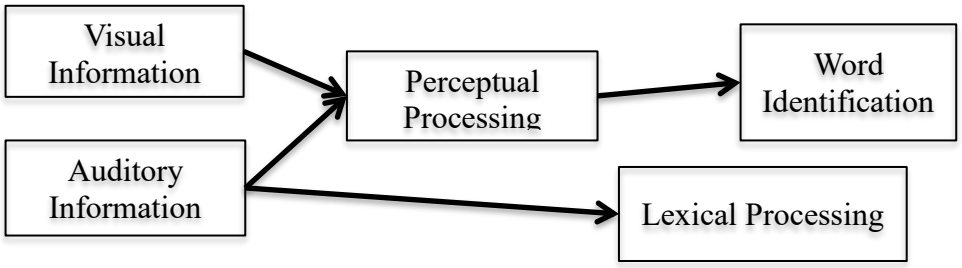


Figure 0.1 shows a schematic of the account of multisensory and lexical processing put forward by Ostrand et al., (2016).

Figure 0.2

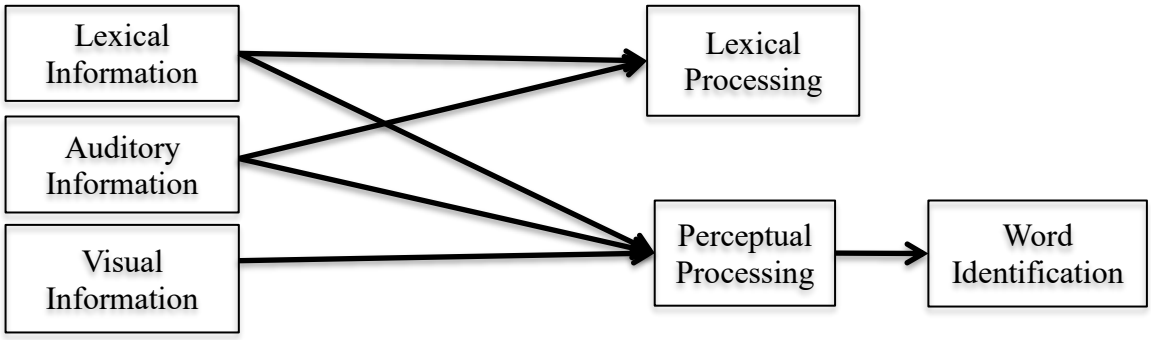


Figure 2 shows a schematic of the account of multisensory and lexical processing put forward by Samuel and Lieblich (2014).

Figure 0.3

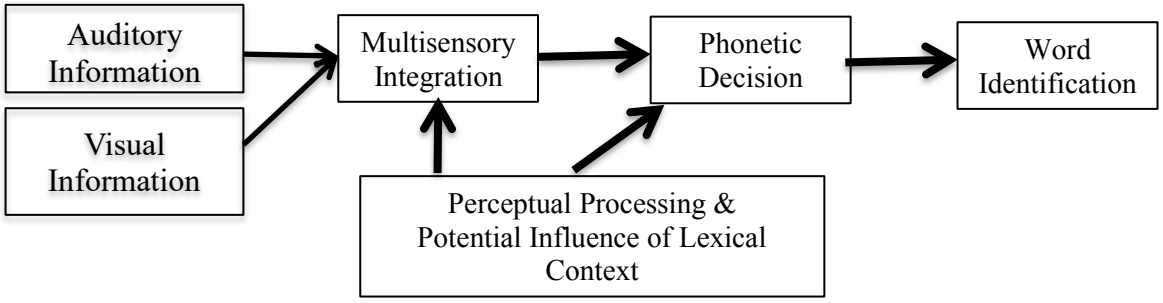


Figure 3 shows a schematic of the account of multisensory and lexical processing put forward by Brancazio (2004).

Chapter 1

Semantic Priming from McGurk Words:

Priming Depends on Perception

Semantic Priming from McGurk Words:

Priming Depends on Perception

Speech perception is inherently multisensory. Seeing the articulations of a talker can enhance perception of auditory speech whether degraded by noise or foreign accent, or even if the speech is clear, but complicated (e.g. Arnold & Hill, 2001; Reisberg, McLean, & Goldfield, 1987; Sumby & Pollack, 1954). Regardless of one's level of hearing, *visual speech perception* is also used during first and second language acquisition, and plays a role in inadvertent speech alignment between interlocutors (e.g. Dias & Rosenblum, 2011; Navarra & Soto-Faraco, 2007; Teinonen, Aslin, Alku, & Csibra, 2008). The multisensory nature of speech is also evidenced by neurophysiological research showing that the speech brain responds to auditory and visual input in remarkably similar ways (for a review, see Rosenblum, Dorsi, & Dias, 2016).

Certainly, the most studied example of multisensory speech perception is the McGurk effect (McGurk & MacDonald, 1976; and for a review see Alsius et al., 2018). The McGurk effect is the finding that if visual speech segments are dubbed onto incongruent auditory segments, the results can be an illusory 'heard' percept that differs from the auditory stimulus. For example, McGurk and MacDonald (1976) report that when auditory 'ba' is dubbed onto a visual 'ga,' perceivers report *hearing* either 'ga' (a visually-dominated perception) or 'da' (a fusion perception). Since its discovery, the McGurk effect has been taken as a hallmark example of audio-visual integration (e.g. Bebko, Schroeder, & Weiss, 2014; Samuel & Lieblich, 2014; Stropahl, Schellhardt, & Debener, 2016; but see Alsius et al., 2018).

The McGurk effect has also motivated much research into how multisensory integration fits into the overall language function. This research has provided both neurophysiological and behavioral data. Much of the neurophysiological work indicates that audio-visual integration occurs early in speech processing (for a review, see Rosenblum et al., 2016). For example, visual speech has been found to produce interactions in the auditory brainstem as early as 11ms following stimulus onset (Musacchia, Sams, Nicol, & Kraus, 2006). This result is consistent with the finding that visual speech produces activity in the auditory cortex (Calvert et al., 1997; Pekkola et al., 2005) as early as 10ms following activation of the visual cortex (Besle et al., 2008). Finally, while likely the result of feedback interactions, visual speech can influence auditory processing in the cochlea as demonstrated by influences on transient-evoked otoacoustic emissions (Namasivayam, Yiu, & Wong, 2015). Collectively, these neurophysiological findings support the assumption that audio-visual integration begins at the earliest stages of speech processing.

These findings are also consistent with behavioral results indicating that multisensory integration occurs very early in the linguistic process. For example, Green and Miller (1985) found that visual speech could produce a McGurk effect that influenced the perception of voice-onset-time (VOT) (See also Brancazio & Miller, 2005; Green & Kuhl, 1989; Sanchez, Miller, & Rosenblum, 2010). As VOT is a pre-phonemic feature of speech perception, this finding suggests that audio-visual integration occurs prior to word, or even word segment, recovery. Similarly, the auditory perception of place of articulation for co-articulated speech is also sensitive to visual speech

information (Fowler, Brown, & Mann, 2000; Green & Norrix, 2001). These findings suggest that multisensory integration begins early, likely before segment features are established, and long before words are identified.

These findings come from work with syllable stimuli, and it is possible that the relative timing of multisensory processing may be different for word stimuli, which carry lexical information (knowledge about the words of the perceivers language). In fact, there is research showing that lexical information may influence multisensory integration. For example, it has been found that McGurk effects are more reliable when they form words as opposed to nonwords (Brancazio, 2004; but see Sams, Manninen, Surakka, Helin, & Kättö, 1998). This finding has been shown to be stronger when the audio-visually discrepant segment occurs in the word final, as opposed to word initial, position (e.g. Barutchu, Crewther, Kiely, Murphy, & Crewther, 2008). Relatedly, this McGurk word bias is stronger when the McGurk word is consistent with the preceding sentence context (e.g. Windmann 2004). These findings suggest that there may be an interactive processing of lexical and multisensory contexts.

What is unclear from these studies is *when* this interaction occurs. Does lexical processing intervene early and influence how unisensory streams are integrated, or are unisensory streams integrated after which the lexical information influences how that integrated percept is categorized (for a discussion, see Brancazio, 2004)? In other words, does lexical processing commence before or after multisensory integration is complete?

To better understand the relative timing of lexical and multisensory processing, Ostrand and her colleagues (Ostrand et., 2016) tested the semantic priming of audio-

visual words. In this task, participants identified audio-only 'target' utterances as words or nonwords, and these targets followed an audio-visual 'prime' word. In such a task, it is generally found that word targets are identified as such faster when they are semantically related to the preceding prime word (e.g. Milberg, Blumstein, & Dworetzky, 1988). Importantly, some of the primes used by Ostrand et al., (2016) were audio-visually congruent (auditory 'bait' + visual 'bait'; auditory 'date' + visual 'date') while others were McGurk-type stimuli (auditory 'bait' + visual 'date' —often perceived as 'date'; see also Brancazio, 2004; Sams et al., 1998; Barutchu et al., 2008; see also MacDonald & McGurk, 1978; McGurk & MacDonald, 1976).

Ostrand and her colleagues' (2016; Experiment 2) found that these McGurk words produced semantic priming more similar to the priming found for words that were audio-visually congruent with the McGurk *auditory* word than for words that were audio-visually congruent with the (ostensibly perceived) McGurk visual word. For instance, auditory 'bait' + visual 'date,' while putatively perceived as 'date,' facilitated the identification of the word semantically related to 'bait' ('worm') but not the word semantically related to 'date' ('time').

From these results Ostrand et al., (2016) concluded that semantic processing preferentially accesses the auditory over the visual—and integrated/perceived—speech information. That is, when both the auditory and visual speech signals contain words, the lexical system processes the auditory information before the visual/integrated information. These results are surprising considering that with McGurk stimuli, participants typically report "hearing" the visually-influenced word. That semantic

processing seems to prefer auditory information even when the auditory word is not the percept suggests that lexical processing occurs prior to the completion of multisensory integration.

The finding that the auditory component of McGurk words determines semantic processing challenges the aforementioned perspective that audio-visual integration occurs early (e.g. Rosenblum et al., 2016). Recall that a large amount of research has suggested that integration occurs at the featural level of speech and that crossmodal influences occur exceptionally early in the neurophysiology of speech processing. In this sense, the findings of Ostrand and her colleagues (2016) are striking and could indicate that a shift in theories of multisensory speech perception is needed. In fact, recent work has attempted to incorporate the results reported by Ostrand et al., (2016) into a coherent theory of multisensory speech perception (e.g. Mitterer & Reinisch, 2017; see also Samuel & Lieblich, 2014; Baart & Samuel, 2015[1]).

Given the potential implications of Ostrand and colleagues' (2016) findings, additional tests of audio-visual speech semantic priming seems warranted. This would seem especially important because their findings are based on semantic priming from McGurk stimuli, and the McGurk effect is known to be quite variable. Research shows that the same McGurk stimuli can produce dramatically different sized effects across subjects within and between experiments (e.g. Brancazio, 2004; Brancazio & Miller, 2005; McGurk & MacDonald 1976; MacDonald & McGurk 1978; for a discussion, see

[1]Samuel & Lieblich, 2014 and Baart & Samuel, 2015 both cite Ostrand, Blumstein, & Morgan (2011) a conference presentation of Experiment 1 from Ostrand et al., (2016).

Alsius et al., 2017). For example, if the McGurk prime stimuli sometimes failed to produce a McGurk effect, that is, if subjects most often perceived the McGurk stimuli as the auditory word, then the semantic priming of the perceived word would be similar to the priming of the auditory word.

Ostrand and colleagues (2016) conducted a pilot study in which perception of the McGurk stimuli was assessed and found that ratings of the integrated McGurk percept were significantly higher than ratings of the auditory signal alone, indicating that participants successfully integrated the audio and visual signals. However, the pilot study asked for goodness ratings of a queried initial consonant (e.g. "how good was the initial D?"), rather than an open-ended input of the participant's perception, and thus participants may not have had the opportunity to rate the consonant that they actually perceived. As a result, if participants sometimes perceived the auditory signal rather than the combined McGurk percept, the semantic priming observed by Ostrand et al., (2016) may have been consistent with the auditory word of the McGurk prime simply because the auditory word was what was most often perceived (with the McGurk effect often 'failing'). If this were the case, then Ostrand et al's (2016) results would actually reflect the *perceived* words, thereby not providing evidence against early integration theories.

Additionally, without knowing how the critical priming words were identified, there may be instances for which a visual influence *does* occur, but not in the way presumed. Consider one of the priming stimuli used by Ostrand and her colleagues (2016): Auditory 'bore' + visual 'gore,' which was assumed to induce a 'gore' perception and was therefore used to prime the target word "blood." However, prior research using

auditory ‘ba’ + visual ‘ga’ syllables indicates that participants often perceive the visually-influenced ‘ga’ less than 30% of the time (e.g. MacDonald & McGurk, 1978). Complicating matters, this same syllable combination is often perceived as the (visually-influenced) *‘da’* more than 60% of the time. Thus, even if subjects identified the stimulus as consistent with the auditory component (‘ba’) infrequently, it is unclear whether the most common percept would be semantically-related to a ‘g’ or ‘d’-initial word (e.g. ‘gore’ or ‘door’). Thus, to understand the relationship between audio-visual integration and semantic priming, analyses examining the specific correspondence of identification and semantic priming responses to McGurk words are clearly necessary.

This, and other considerations, motivated a follow-up study that replicated the original design of Ostrand et al., (2016) with two chief differences. First, a single set of participants performed both the priming test and an identification test of our McGurk stimuli. This was instituted to establish that for the tested group of participants, the visual influence was large enough, and occurred in the predicted way, to have the potential to induce semantic priming based on the perceived words. Inclusion of a McGurk identification task also allowed us to correlate the observed semantic priming responses with the participants’ perception of the stimuli. Second, to add to the stability of the stimuli, we chose words with McGurk segment combinations that are known to induce very large visual influences. For these purposes, we restricted our McGurk stimuli to auditory ‘b’ and visual ‘v’ -initial words. Prior research has demonstrated that the auditory ‘ba’ visual ‘va’ McGurk combination is very reliable (e.g. 98% ‘va’ perceptions; Rosenblum & Saldaña, 1992).

If semantic analysis is based on the perceived, rather than auditory component of audio-visual words, then these stronger McGurk segments should induce a priming effect based on the visual component. Further, the likelihood of the visual/perceptual-based priming should correspond to the observed strength of the McGurk effect for each word combination. If, on the other hand, semantic analysis is based on the non-integrated auditory component of a McGurk stimulus, then priming should correspond to the auditory words, despite evidence for a strong McGurk effect in the identification results.

Main Experiment

Study 1 replicated the design of the semantic priming task used by Ostrand et al., (2016) but here we changed the stimuli used in that task with the goal to foster consistent McGurk-visual perceptions. If these stimuli show semantic priming to the visual word, then it is likely that semantic processing is of perceived lexical information. To further understand the relationship between McGurk perception and semantic priming, subjects were asked to perform an identification task of the McGurk stimuli, following the semantic priming paradigm. If lexical processing is preferentially related to the auditory word components, as opposed to perceived words, then this identification measure should not be related to semantic priming. If, however, semantic processing responds to the perceived (and integrated word), then there should be a correspondence between identification judgments and semantic priming.

Method

Participants

Participants were 119 native English speakers from the University of California, Riverside. All participants reported having normal hearing and vision. All participants were compensated with either course credit or $10.00 cash.

Materials

The stimuli were audio-visual word primes followed by auditory-only word or nonword targets. Following the method of Ostrand and her colleagues' (2016) second experiment, a 50ms interval separated the offset of the audio-visual prime and the onset of the auditory-only target. All stimuli were produced in a single recording session by a male, native monolingual English speaker. The speaker had lived in Southern California for approximately 4 years prior to recording.

Our central question concerned the semantic priming produced by McGurk stimuli. Our McGurk primes consisted of pairs of English words differing only in their initial consonant. For the critical stimuli, we used only words that began with either 'b' or 'v' (e.g. auditory 'bale' + visual 'veil'). The motivation for this was two-fold. First we wanted to be confident that McGurk primes rarely resulted in the participants perceiving the auditory word (i.e. auditory 'bale' + visual 'veil' perceived as 'bale'; a 'McGurk-auditory'). The key question addressed by this experiment, and by Ostrand and her colleagues (2016), is if lexical processing is related to the auditory stimulus, independent of perception. If a McGurk prime frequently produces a percept of the auditory stimulus (e.g. a McGurk-auditory perception) then answering this question will be difficult.

Second, we wanted to be confident that when the McGurk effect does occur, participants will perceive the predicted visually-dominated word (auditory 'bale' + visual

'veil' = perceived 'veil'). Past research has shown that the auditory 'ba' visual 'va' combination produces a high frequency of visually-dominated percepts (e.g. ~98%; Rosenblum & Saldaña, 1992) and was thus ideal for our design.

We identified 24 /b/-initial—/v/-initial minimal word pairs to be used as McGurk stimuli (Table 1.1). A pilot study consisting of 27 participants was conducted to test the strength of the visual influence of these word combinations. Using an open response identification task it was found that these 24 audio-B visual-V McGurk words produced visually-dominated responses 74% of the time. While this average is notably smaller than the 'v-b' visual dominance reported in other studies (~98%; Rosenblum and Saldaña, 1992), it should be noted that those previous studies tested perception of syllable stimuli in a two-alternative force-choice task (see also Brancazio, 2004).

Each of our McGurk primes was assigned a semantically related target. Following the procedure of Ostrand et al., (2016), related targets were selected from the University of South Florida Free Association Norms database (Nelson, McEvoy, & Schreiber, 1998) and the Edinburgh Associative Thesaurus (Kiss, Armstrong, Milroy, & Piper, 1973). We also considered data collected from a norming study from students at UC Riverside (again following the procedure for Ostrand et al., 2016). From these three sources of information, we chose the targets that optimized semantic relatedness, but reduced phonological similarity between primes and targets (see Ostrand et al., 2016). This choice was made because prior work has found that visual speech stimuli can phonologically prime audio-only speech targets (e.g. Fort et al., 2013). The complete list of words used

as the McGurk auditory and visual stimuli for the primes and the targets related and unrelated to these words is presented in Table 1.1.

Across participants, each of these primes was paired with four targets: a target related to the visual word, a target unrelated to the visual word, a target related to the auditory word, and target unrelated to auditory word (Ostrand et al., 2016). Each semantically related target was presented as an unrelated target for another prime and thus acted as its own control (Ostrand et al., 2016). Nonword targets were replicated from Ostrand and her colleagues' (2016) second experiment. Additionally, several filler primes were included; these items were also taken from Ostrand and her colleagues' (2016) second experiment. All auditory stimuli were presented through sound insulated headphones (Ostrand et al., 2016) at an average of 70db.

Procedure

The experiment procedure contained two parts. First, participants preformed a lexical decision task that assessed the semantic priming of McGurk and audio-visually congruent word stimuli. Second, participants performed an identification task that assessed their perceptions of the McGurk and audio-visual congruent words used as primes in the lexical decision task, as well as the audio-only versions of those stimuli.

During the lexical decision task, participants were instructed to watch and listen to the audio-visual prime word and then listen to the audio-only target (Ostrand et al., 2016). The participants were instructed to indicate if the target was a word or nonword by pressing one of the two labeled buttons on a button box. Participants were instructed to

respond as quickly and accurately as possible (Ostrand et al., 2016). The word/nonword button assignment was counter-balanced across participants (Ostrand et al., 2016).

For the lexical decision task, 1/3 of the b/v initial primes were presented as McGurk stimuli (e.g. auditory 'bale' + video 'veil'). The remaining 2/3s were presented audio-visual congruently, and were equally divided between two types of audio-visual congruent stimuli (Ostrand et al., 2016). The first of these were b-initial words (*b-congruent*) made up of the auditory components of the McGurk stimuli (e.g. auditory 'bale' + video 'bale'). The second type of audio-visual congruent stimuli was v-initial words (*v-congruent*) made up of the visual components of the McGurk stimuli (e.g. auditory 'veil' + video 'veil'). The items chosen as *McGurk*, *b-congruent*, and *v-congruent*, were counter-balanced across participants (Ostrand et al., 2016). The condition design of the experiment is portrayed in Appendix A.

Half of all trials included nonword targets. In order to test semantic priming of the critical *McGurk*, *b-congruent*, and *v-congruent* tokens, the 24 McGurk primes were *only* used for word trials, and filler items were used for the nonword target trials (Ostrand et al., 2016). These filler words were the same as those used by Ostrand and her colleagues' (2016) (Experiment 2) and included an array of words with initial consonants other than 'b-v' combinations (e.g. audio 'pad' + video 'tad'; audio 'mine' + video 'nine'). To reduce the potential of participants statistically learning that 'b' and 'v' initial words preceded word targets, we also included 12 non-b/v initial filler primes with word targets also recorded from the same speaker.

To reduce the potential of participants learning that McGurk items were more likely to precede word targets, half of the nonword target trials were preceded by non-b/v initial McGurk primes (Ostrand et., 2016). Finally, to reduce the possibility that participants might learn that the b/v initial McGurk primes consistently lead to word targets, four of the 12 filler primes preceding word targets were also non-b/v initial McGurk words (Ostrand et al., 2016).

Thus, for each subject, the lexical decision task included 12 McGurk primes (8 critical; 4 filler) with word targets and 12 McGurk primes with nonword targets (all filler). Additionally, the task included 24 congruent prime-words with word targets (8 b-congruent; 8 v-congruent; 8 filler) and 24 congruent primes with nonword targets. This corresponded to 72 total trials for each subject. Subjects were given one self-timed break administered between trials 36 and 37 (Ostrand et al., 2016).

The 24 critical prime words were distributed into 12 different conditions for each participant (Ostrand et al., 2016). These twelve conditions included: targets *related* and *unrelated* to the McGurk visual word, and targets *related* and *unrelated* to the McGurk auditory words (Ostrand et al., 2016). For the primes used for each subject, the unrelated words were items that actually served as related targets for primes presented to different subjects. In this sense, these words acted as their own controls (Ostrand et al., 2016).

These four conditions were repeated for primes that were audio-visually congruent with the McGurk visual ‘v’ word (v-congruent) and primes that were audio-visually congruent with the McGurk auditory ‘b’ word (b-congruent). Each critical prime

was only presented once to each participant, and which primes were placed in which condition was counterbalanced across participants (Ostrand et al., 2016).

To summarize, each subject was presented eight of the critical b-v incongruent priming McGurk items (which of the eight were presented was counterbalanced across subjects). Two of these McGurk items were followed by targets related to the (b-word) audio component of the McGurk stimulus, and two were followed by targets related to the (v-word) visual component. The remaining four critical b-v McGurk primes were followed by target words unrelated to the prime words (and served as comparison trials). All of the remaining 72 trials included audio-visual congruent control items (16) and unscored filler items (48).

During the experiment, participants were seated approximately 30 inches from the computer screen. Each trial began with a white '*' fixation point presented for 600ms on a black background. Immediately following the fixation point, the face of the talker appeared and articulated the prime word (the fixation point was aligned with the center of the talker's lips). After the articulation of the prime word, the screen went blank (Ostrand et al., 2016). Fifty milliseconds following the acoustic offset of the prime word, the target word was presented through the same headphones, without any accompanying visual stimulus on the screen (Ostrand et al., 2016). The trial ended when the participant pressed either the 'Word' or 'Nonword' buttons on a button box (Ostrand et al., 2016).

Following the completion of the lexical decision task, participants started the identification task. A programing error resulted in identification data not being collected for one of the 24 McGurk items (audio 'buy' + visual 'vie'). Thus, participants were

presented with 23 McGurk items along with the corresponding 48 audio-visual congruent items that were used as primes for the lexical decision. In addition, they were also presented the 48 audio-alone words which comprised the audio-visual items. Items were blocked by audio-visual vs. audio-alone stimulus type and randomized within blocks. Participants were instructed to attend to each utterance and to use the keyboard to type the word they *heard* the talker say (e.g. Alsius et al., 2018). Participants were not informed that the stimuli were the same items from the lexical decision task, and were not informed that the items would all be words. Participants were allowed to view their responses as they typed them and were instructed to correct any errors or typographic mistakes before proceeding to the next trial. As in the priming task, each audio-visual trial included a fixation point at the location of the talker's lips that was present for 600ms immediately preceding the appearance of the talker's face.

Results

Semantic Priming Reaction Times

Only reaction times from trials that included one of our 24 critical primes (the McGurk words) and their 48 congruent counterparts were analyzed (Ostrand et al., 2016). Responses that that were incorrect (6.4%), occurred before the target word offset (7.1%), or that were more than two standard deviations from the condition mean reaction times (3.2%) were excluded from the analysis (Ostrand et al., 2016).

Following Ostrand and her colleagues (2016), we submitted these reaction times to both a subject analysis and an item analysis. Each analysis began with an omnibus ANOVA consisting of the following factors: 2 Relatedness (Related vs. Unrelated) x 2

Target (associated with: Visual word vs. Auditory word) x 3 Prime (McGurk, v-congruent, or b-congruent). Condition means for the subject analysis are displayed in Figure 1.1a.

Consistent with what is reported by Ostrand et al., (2016) neither ANOVA found significant main effects of Prime. As was found by Ostrand et al., (2016), both subject (F_1) and item (F_2) tests returned significant effects of relatedness; $F_1(1, 94) = 21.193$, $p < .001$, $\eta^2_p = .184$ and $F_2(1, 23) = 4.647$, $p = .042$, $\eta^2_p = .168$, indicating that across conditions, targets were identified as words faster when they were semantically related to the preceding prime (M_1: 329 vs. 371).

The subject, but not the item, analysis showed a significant main effect of target type; $F_1(1, 94) = 16.762$, $p < .001$, $\eta^2_p = .151$ versus $F_2(1, 23) = 1.879$, $p = .184$, $\eta^2_p = .076$. The effect of target in the subjects' analysis indicates that both related and unrelated targets associated with the visual word (e.g. 'veil'→'wedding') were identified faster than targets both related and unrelated with the auditory word (e.g. 'bale'→'hay'; M_1: 361 vs. 340). While we refrain from interpreting the null effect of the item analysis, it is worth noting three observations. First, despite failing to produce a significant effect, the means from the item analysis show a similar pattern to the means from the subject analysis, with the McGurk visual associates being identified faster than the McGurk auditory associates (M_2: 360 vs. 337). Second, Ostrand and her colleagues (2016) also report a significant effect of target for their subject analysis, but not their item analysis (see pg. 102).

Third, while both our results and the results of Ostrand et al., (2016) show significant effects of the same factors, the patterns of the condition means indicates that our effects have a different locus than the effects of Ostrand et al., (2016). Specifically, while Ostrand and her colleagues' (2016) subject effect (and item trend) was driven by faster responses to the stimuli's *auditory* associated targets, ours were driven by faster responses to the *visually* associated targets (see Figure 1.1).

Both our subject and item analyses found significant two-way interactions between Relatedness and Target association; $F_1(1, 94)= 4.450$, $p = .038$, $\eta^2_p = .045$, $F_2(1, 23)= 5.571$, $p = .027$, $\eta^2_p = .195$. These interactions indicate that the priming effect for auditory associated targets was different than the priming effect for visual associated targets. None of the remaining two-way interactions for the omnibus tests were significant for either the subject or item analyses.

The most important effect returned by the omnibus test is the three-way interaction between Relatedness, Target association, and Prime stimulus (Ostrand et al., 2016). Both the subject ($F_1[2, 188]= 9.965$, $p < .001$, $\eta^2_p = .096$) and the item ($F_2[2, 46]= 9.115$, $p < .001$, $\eta^2_p = .284$) analyses revealed that this interaction was significant. This interaction is portrayed in Figure 1.1a. This interaction indicates whether a prime produced semantic priming for auditory associated targets or the visual associated targets depended on whether the prime was a McGurk stimulus, v-congruent stimulus (associated with the visual McGurk component), or b-congruent stimulus (and associated with the auditory McGurk component). Importantly, it is this interaction that allowed

Ostrand and her colleagues (2016) to conclude that the McGurk prime induced responses more similar to the auditory than visual component of the stimulus.

However, the pattern of results portrayed in Figure 1.1a tells a different story. This figure shows that targets related to the visual channel were identified faster than other targets, both for the McGurk and v-congruent primes. In contrast, when the prime was b-congruent (and consistent with the audio component of the McGurk stimulus), targets related to the auditory channel were identified fastest. Thus, as can be seen in Figure 1.1a, priming responses to the McGurk stimulus were more similar to the v-congruent than b-congruent stimuli.

As the prime stimulus factor had three levels, additional analyses were needed to determine the true locus of the interaction, and whether it indicates that the effect is driven by the difference between the b-congruent (*auditory*) condition relative to the McGurk and v-congruent conditions, as suggested by the plots.

Post-Hoc Tests. To identify the locus of the interaction, we computed ANOVAs examining each pairing of 2 of the 3 Prime conditions in 2 (Related) x 2 (Target) x 2 (Prime) ANOVAs (Ostrand et al., 2016). Again these analyses were computed by subjects (F_1) and by items (F_2). The results of these analyses are shown in Table 1.2. The most important results of these analyses are the three-way interactions that indicate that the priming effect for auditory associated and visual associated targets is modulated by the prime stimulus. As can be seen in Table 1.2, this three-way interaction is present when comparing the McGurk and b-congruent primes ($F_1[1, 98]= 11.166$, $p = .001$, $\eta^2_p = .102$; $F_2[1, 23]= 7.402$, $p = .012$, $\eta^2_p = .243$) and when comparing the b-congruent and v-

congruent primes ($F_1[1, 95]= 17.044$, $p < .001$, $\eta^2_p = .152$; $F_2[1, 23]= 22.905$, $p < .001$, $\eta^2_p = .499$). This interaction was not significant when comparing the McGurk and v-congruent primes ($F_1[1, 98]= 1.306$, $p = .256$, $\eta^2_p = .013$; $F_2[1, 23]= 1.615$, $p = .216$, $\eta^2_p = .066$).

Together these results indicate the priming effect of auditory associated and visual associated targets is modulated by the difference in the priming effects between the McGurk and b-congruent (audio) primes and between the b-congruent and v-congruent primes. Put differently, the McGurk and v-congruent (visual-related) primes induce similar effects, both of which are different from those induced by b-congruent (audio) primes. This supports the interpretation that semantic priming with the McGurk stimuli was related to the *visible*—and possibly perceived—word component rather than the auditory component.

Our results contrast markedly with the results reported by Ostrand and her colleagues (2016), who found that it was their McGurk and b-congruent (*audio*-related) primes which induced similar responses.

Identification Task Responses

The question naturally arises of why do our current results contrast so dramatically with the results of Ostrand and her colleagues (2016)? One hypothesis is that the differences are attributable to the relative strength of the McGurk effects in the two studies. As stated, the McGurk stimulus identification results were not available from the Ostrand et al., (2016) study. However, the hypothesis can be indirectly examined by evaluating the identification data collected in the current study.

In analyzing these identification responses we had to consider how best to measure the McGurk effect, based on our free-response task. The operational definition of the McGurk effect varies in the literature, with some researchers defining only identifications that differ from both the auditory *and* visual stimulus as the McGurk effect (e.g. van Wassenhove, Grant, & Poeppel, 2007; Magnotti & Beauchamp, 2015) while others define the effect as any instance in which the visual stimulus changes the perception of the auditory stimulus (e.g. Brancazio, 2004; Rosenblum & Saldana, 1992; see also Alsius et al., 2018). Neither of these definitions would be sufficient to analyze the results of the present study as we were concerned with the correspondence between the influence of *specific* priming words and the pattern of McGurk identifications. For these reasons, we chose to calculate two separate identification scores for each of our McGurk items: the percentage of auditory word responses and visual word responses. Note that because we used an open-response task, subjects were allowed to provide responses that corresponded to neither the auditory or visual word. However, because these types of responses did not correspond to the auditory or visual prime components, they were not used in this analysis.

McGurk Rates. We analyzed our data by tabulating participant responses that began with the letter 'b' and those that began with the letter 'v.' A technical problem resulted in the data from 21 participants not being collected from the identification task. Accordingly, data from the remaining 98 participants were used to analyze stimulus identifications. We found that our stimuli produced robust McGurk effects with visually-based responses provided on 68.3% of trials and auditory-based responses on 19.1% of

trials. These data are similar to those found for our pilot study and the strength of the effect is comparable to other studies that have used word stimuli and free-response tasks (see above). The data also showed a wide range in the proportion of visually-based responses across the different items (e.g. 86.1% for 'bowel'—19.4% for 'beer') as well as a wide range in the proportion of auditory-based responses (e.g. 48.5% for 'bury' and 4.2% for 'bale'). A number of factors likely account for these differences in effect strength across items including word frequency and neighborhood density characteristics (see also Brancazio, 2004; Barutchu et al., 2008; Chapter 3). A summary of the identification responses for all our items is provided in Table 1.3.

Preparing McGurk Identification Rates for Semantic Priming Analysis. To address our hypothesis that the strength of the visual influence on speech perception modulates the semantic priming of McGurk words, we included the McGurk rates as a covariate in the item analysis of our reaction time data. For this analysis we converted our identification data into McGurk *identification-differentials* by subtracting the auditory-based response rate from the visually-based response rate for each McGurk prime. For example, the McGurk stimulus auditory 'bane' + video 'vein' was perceived as 'bane' (McGurk-auditory) 12.7% of the time and as 'vein' (McGurk-visual) 77.7% of the time, and thus produced a McGurk identification-differential of 65%. In this way, this identification-differential conveyed the relative frequency of the two outcomes that could be expected to influence semantic priming. This difference score also isolated our McGurk measurement from the irrelevant nonvisual-McGurk responses, removing a substantial source of variability from the analysis.

For each McGurk item, the identification-differential was calculated based only on identification responses from the participants who also provided priming-task reaction times containing that particular McGurk item. Recall that during the semantic priming task, each participant was presented only *eight* critical incongruent McGurk words. Thus, the identification-differential score for each item only included identification data from the specific participants who had been presented that word in McGurk format during the semantic-priming task. Recall also that the reaction times submitted to the semantic priming analyses were subject to exclusion criteria (see above). Thus, if a participant's reaction time value for a McGurk stimulus in the priming experiment was excluded from the analysis, their corresponding identification response for that stimulus was also excluded from the identification-differentials calculation.

Interaction of McGurk Scores and Reaction Times. To infer semantic priming from our design, we needed to compare the reaction times from twelve conditions (see methods). As stated, each participant only received each McGurk item in one of these conditions. This made it impractical to use McGurk rates in the subject analysis (F_1). Instead we included the identification-differential for each item and included it in the *item* analysis (F_2). This ANCOVA retained the significant three-way interaction between relatedness, target, and prime ($F_2[2, 42]= 4.687$, $p = .015$, $\eta^2_p = .182$), indicating that the pattern of semantic priming depended on the stimulus type of the prime. More importantly, this analysis also returned a four-way interaction between those factors and the identification-differential ($F_2[2, 42]= 3.614$, $p = .036$, $\eta^2_p = .147$). This four-way

interaction suggests that the pattern of semantic priming was modulated by the strength of the visual effect on perception.

Correspondence Between Semantic Priming and Perception. To characterize the relationship between perception and semantic priming, we conducted a correlation test between priming scores and McGurk rates. Semantic priming was calculated through the relationship between two sets of reaction times: those derived from targets related to a prime and those derived from targets unrelated to a prime. For a McGurk prime, priming must be calculated for both targets related/unrelated to the auditory *and* visual components of the prime. Thus the four-way interaction in the ANCOVA was not driven by any single set of reaction times, but from the relationship across four sets of reaction times. Just as the McGurk effect needed to be measured in a way that conveyed both the rate of auditory and the rate visual consistent identifications, we needed to measure semantic priming in a way that conveys priming from both the auditory and visual components of a McGurk prime.

For this reason, we calculated *priming-differential* scores from our lexical decision task reaction times. These priming-differential scores were calculated from only reaction times that were collected from McGurk trials from the lexical decision task. To calculate the priming-differentials these reaction times were divided into four groups: reaction times to targets (1) related and (2) unrelated to the McGurk audio word, (3) related and (4) unrelated to the McGurk visual word. Using these four groups of reaction times, the priming-differential scores for each McGurk word item were calculated in three steps. First, for each McGurk word prime, the reaction times for the target related to

the McGurk auditory word were subtracted from the reaction times for the target unrelated to the McGurk auditory word. This step generated an *auditory priming score* for each McGurk prime. Larger auditory priming scores indicate that reaction times to targets related to the McGurk auditory prime were shorter than the reaction times to targets unrelated the McGurk auditory prime. Thus, positive auditory priming scores indicate that the auditory component of a McGurk stimulus semantically primed the identification of the targets. This process was then repeated for reaction times to targets related and unrelated to the McGurk visual words, forming *visual priming scores*. Finally, for each McGurk stimulus, the auditory priming score was subtracted from the visual priming score, forming the *priming-differential*. This metric provided an estimate, for each McGurk word, that indicated how likely it was to produce priming to its audio component relative to priming to its video component.

A correlation test was then conducted on the *identification-differential* and *priming differential* scores for each McGurk item. The correlation between the two differentials is shown in Figure 1.2 and was found to be $r = .388$, $p = .034$ (1-tailed) for the 23 McGurk items tested. This correlation indicates a relationship in which items that were more likely to produce McGurk-visual perceptions were also more likely to produce visual priming effects. This finding is consistent with our hypothesis that semantic priming is related to the *perception* (identification) of the prime.

Given the multiple constraints that shaped the primes and targets used in this study (i.e. b-initial/v-initial minimal word pairs for primes; targets with distinct semantic relationships between auditory & visual words; targets with minimal phonological

similarity to primes, etc.) there are many sources of variability affecting the semantic priming scores. Finding that a significant portion of this variability, even if only a small amount, is accounted for by the strength of the McGurk effect demonstrates how influential perception is in semantic priming. Finally, it should be noted that this conclusion does not preclude the possibility that semantic priming may sometimes be related to the auditory stimulus. We argue, instead, that semantic priming will be consistent with the auditory stimulus when the auditory stimulus is what is perceived.

Follow-up Analysis: Cross Lab Investigation

A follow-up analysis sought to determine if the effects found in our experiment, that the semantic priming of McGurk stimuli was related to the identification of those stimuli, could also account for the results of Ostrand and her colleagues' (2016) Experiment 2. This study made use of results of semantic priming provided in the Ostrand et al., (2016) report, and also of unpublished identification data for the stimuli used in that study but collected after the Ostrand et al., (2016) report was published (and using different subjects). From these data, the rates of visual-based and auditory-based identifications for each McGurk stimulus of the Ostrand et al., (2016) study were used to illuminate the relationship between the McGurk effect and semantic priming.

Method

Participants

265 students from the University of California, San Diego participated in this experiment for course credit. All participants were native English speakers and reported having normal hearing and vision.

Materials

The stimuli in this experiment were produced by a 24 year old female native English speaker from Rhode Island. The speaker produced 72 words that were used to generate the 36 McGurk stimuli that were used in Ostrand et al., (2016). These words were minimal pairs, always differing in only the initial consonant. McGurk words included; audio 'b' + visual 'd', audio 'p' + visual 't', audio 'p' + visual 'k', audio 'b' + visual 'g', and audio 'm' + visual 'n' pairings. Further details of these stimuli can be found in the original paper (Ostrand et al., 2016).

Procedure

Participants wore sound insulated headphones while observing the speaker say each of the 36 McGurk words on a computer screen in front of them. Participants were instructed to watch and listen to each word carefully. Participants were instructed use the keyboard to report the initial consonant from the start of each McGurk word.

Results

McGurk Rates

Responses to the McGurk stimuli used in the Ostrand et al., (2016) study were tabulated for proportion of visually and auditory-based responses (See Figure 1.3; see also Table 1.6 for item means). The visually-based response rate for these stimuli was 39.7%. This rate is significantly smaller ($t[57] = -4.606$, $p < .001$) than the visually-based response rate found for the stimuli used in our own experiment ($M = 68.3\%$). Relatedly, the auditory-based response rate for stimuli in the Ostrand et al., (2016) study was 35.8% which is significantly larger ($t[57] = 3.987$, $p < .001$) than the McGurk-auditory rate

found in Study 1 above (M = 19.1%). Clearly, these data suggest that the data used by Ostrand et al., (2016) failed to produce as strong a McGurk effect as did the stimuli used in our Study 1. In addition, it seems that those stimuli produced a substantial number of responses (24.5%) that corresponded to *neither* the auditory or visual component.

While based on a different group of subjects (tested at a later date), this analysis could indicate that participants in the Ostrand et al., (2016) Experiment 2 were less likely to experience the predicted McGurk effect than were the participants in our experiment. Potentially then, participants in the Ostrand et al., (2016) study may have often perceived the auditory word of the McGurk stimuli during the semantic priming task. Recall that Ostrand et al., (2016) found semantic priming consistent with the auditory component of McGurk stimuli. These new data raise the possibility that this finding may have actually been driven by the *perception* of the McGurk stimuli—which was often consistent with the auditory component—rather than by the *unintegrated* auditory channel, as such.

Correlations

The correspondence between McGurk effect identifications and semantic priming was next calculated with the data from Ostrand et al., (2016) stimuli. For these purposes, a correlation test was conducted for McGurk items using the previously described *identification-differential* scores and *priming-differential* scores. It should be noted that here, the priming-differential scores were calculated based on the reaction time values of Ostrand et al's (2016) Experiment 2, and identification-differential scores obtained at a much later date. Thus, unlike the correlation conducted for our own experiment (above), the correlation based on the stimuli used by Ostrand et al., (2016) was calculated *across*

two different subject groups. Because of the known inter-subject differences in McGurk effect responses (for a review, see Strand et al., 2014), it might be expected that this correlation would not be as strong as when the same subjects are used for both McGurk priming and identification tasks.

The correlation between the *identification-differentials* and p*riming-differentials* for Ostrand et al's (2016) McGurk stimuli was found to be $r = .170$, $p = .165$ (n=35[2]). While this correlation was not significant, it is interesting to note that: a) an r value of .17 is considered to be between a weak and moderate effect (Cohen, 1992); and b) the correlation is in the same direction as the correlation reported for our own study above. This is interesting because this outcome, while marginal, contrasts with the initial interpretation of the semantic priming reported by Ostrand et al., (2016). Accordingly, this correlation between the semantic priming reported by Ostrand et al., (2016) and these newly acquired identification rates tentatively suggests that, in contrast to their initial conclusions, Ostrand and her colleagues (2016) may have observed semantic priming related to the perceived word rather than the auditory stimulus.

Cross-Study Comparison

A notable difference between this correlation and the correlation of our own study is that this correlation is much smaller. As stated, however, this correlation used identification-differentials that were calculated from a separate set of participants from those who provided the reaction time data for the priming-differentials. Potentially, this between subject group calculation is responsible for the reduced magnitude of the current

[2] Reaction times for the audio "part" + visual "tart" prime with an unrelated target from Ostrand et al., was not available and thus no priming differential could be calculated.

correlation. To test this possibility, we re-calculated the McGurk-differential for our own study; this time only including identification data from participants whose reaction times were excluded from the priming differential.

Recall that in our own study, the identification-differential scores for the 23 critical McGurk stimuli were calculated based only on identification responses for participants who had also provided reaction times for those *particular items* in the lexical decision task. Because each subject only responded to only eight of the critical McGurk tokens in the lexical decision task, only a subset of the participants received any given audio-visual item in incongruent McGurk (as opposed to congruent) format. Thus, for the current analysis, identification-differential scores for any McGurk item were calculated based on participants *who did not* provide critical reaction times to those McGurk items during the lexical decision task. It was thought that this cross-subject analysis would be more similar to the analysis calculated across subjects for the data collected on the Ostrand et al., (2016) stimuli.

We found that calculating the correlation for our own data using these new between subject identification-differential scores did indeed reduce the strength, but not the direction, of the correlation reported in our own experiment (r = .166, p = .225 [1 tailed], n=23). Potentially, this weakened effect is a direct result of calculating the correlation across two measures from two different subject groups. Interestingly, this between subject correlation is very similar to the between subject correlation based on the stimuli used by Ostrand and her colleagues (2016; r = .170, p = .165, n=35). In fact, a z-test comparing these correlations (Diedenhofen & Musch, 2015) revealed that they were

not significantly different from each other, $z = .014$, $p = .5$. This result is suggestive that for Ostrand et al's (2016) stimuli, the weakened nature of the McGurk priming x identification correlation may also be a result of different subjects being used to derive the two measures. A final correlation analysis was conducted in which the data were pooled for the 58 total critical stimuli of both the current, and Ostrand et al., (2016), cross-subject comparisons. This analysis revealed a significant correlation, $r = .273$, $p = .018$ ($n = 58$) (See also figure 1.4). While this pooled analysis must be interpreted cautiously (e.g. the two experiments were not conducted together and used different stimuli generated by different talkers), it is consistent with the results of our own experiment showing that the strength of the McGurk effect is predictive of whether the visual or auditory component of the stimulus is the stronger prime.

General Discussion

The purpose of this investigation was to provide a rigorous test of the hypothesis that auditory speech information has a privileged status over visual, or audio-visually integrated information during semantic processing. Ostrand and her colleagues (2016) reported that with McGurk words, semantic priming was consistent with the auditory component. This finding was surprising in suggesting that lexical processing may commence prior to, or at least concurrent with, multisensory integration. As stated, this conclusion seemed at odds with much of the multisensory speech data suggesting very early integration of the streams – likely at the featural level of linguistic processing (e.g. see Rosenblum et al., for a review).

However, the conclusion offered by Ostrand, and her colleagues was dependent on the assumption that the McGurk stimuli used in their study consistently produced visually-based perceptions. As Ostrand et al., (2016) lacked an identification assessment of their McGurk stimuli, we set out to replicate their original experiment with an identification measure, and implement stimuli known to induce strong visually-influenced responses.

Based on these changes, we found evidence that semantic priming of audio-visual incongruent speech more closely follows the visual word than the non-perceived, auditory component. We further found that the degree of semantic priming for a visual stimulus was correlated with the rate of visually-based identifications for that McGurk stimulus, suggesting that semantic priming followed the perceived (and integrated) word. Importantly, this interpretation allows for semantic priming to sometimes be based on the auditory component, specifically when the McGurk effect fails, and perception *is of* the auditory component.

This interpretation may help account for the findings reported by Ostrand and her colleagues (2016). In fact, the new identification results of their stimuli reported above are consistent with this interpretation. These new identification results show that Ostrand et al's (2016) stimuli are identified as the visual-based words and auditory-based words at similar rates (39.7% and 35.6%, respectively). Additional analyses of their stimuli show a trend that an item's identified visual influence is marginally related to whether it induced a visually-based priming effect. Thus, it is possible that Ostrand and her colleagues' (2016) finding that the auditory channel of McGurk stimuli often drove semantic priming

is a result of participants often *perceiving* those stimuli as consistent with the auditory channel.

Of course, a number of factors distinguished the current study from that of Ostrand and her colleagues (2016) including the McGurk segment combinations used (b/v *vs.* b/d, b/g, p/t, p/k, m/n), the words tested, and the talker used to create the stimuli. It is possible that these differences induced a different processing strategy such that the auditory *rather than* visual/perceptual component provided the basis for semantic priming. It could be, for example, that only for b/v combinations does integration precede semantic analysis, and that for all other combinations (e.g. b/d, b/g, p/t, p/k, m/n), semantic analysis occurs first. However, the notion that a completely different processing strategy is used for different syllable combinations would certainly be the less parsimonious explanation, particularly in light of the similar correlations across stimulus sets. This would seem especially true given the overwhelming support for early multisensory integration discussed above (for a review, see Rosenblum et al., 2016). Instead, we argue that the most likely explanation for the different priming results between the current, and Ostrand et al., (2016) studies, simply lies with the degree to which the expected visual influence on perception (the McGurk effect) actually occurred.

Visual Dominance vs. Fusion Integration in the McGurk Effect

It is worth noting that both the current study and that of Ostrand et al., (2016; Experiment 2) used target words that related either to the auditory or visual component of the McGurk priming stimulus. Thus, neither study tested the priming of words beginning with *fused* segments. As stated, the choice to use visual dominant McGurk stimuli in the

current study was based on: a) attempting to induce the strongest possible McGurk influence; and b) to limit the complexity of prime/target/foil design which was already cumbersome.

Arguably, however, not using *fused* segments could constrain the conclusions that can be made about audio-visual *integration*, as such, and its relation to semantic priming. However, as has been argued (e.g. Alsius et al., 2017), it is unlikely that visual-dominance vs. fusion McGurk effects tap into different processing strategies, especially if subjects are instructed to base responses on what they "hear." This notion is supported by similar patterns of behavioral and neurophysiological results for both classes of stimuli (e.g. Burnham & Dodd, 2004; Jordan, McCotter, & Thomas, 2000; Saldaña & Rosenblum, 1994; Shahin et al., 2018). Thus, we would predict that as for the current stimuli, fusion stimuli would induce semantic priming related to the perceived word. Experiments are planned in our laboratory to test this prediction.

Additional future work should also consider methods of addressing the key assertion made by Ostrand et al., (2016) that time determines the lexical processing of audio-visual speech. While our results demonstrate that semantic priming is consistent with the perceived word of a McGurk stimulus, they do not address the time course of those lexical processes. It is, for example, possible that the integration of the incongruent audio-visual information did delay semantic processing. The most direct way to address this question would be to manipulate the time between the presentation of the prime and target stimuli (e.g. see Neely, 1977). However, doing this will be complicated by the fact that the time between primes and targets is confounded with the duration of the prime and

target utterances. This difficulty is further compounded by the fact that each word utterance has a different recognition-point (e.g. Marslen-Wilson, 1987).

One way that might be helpful in addressing this question could be using neurophysiological measures. The N400 ERP is known to be sensitive to semantic processing (e.g. Delaney-Busch et al., 2019) and there are recent reports that analysis of EEG data can differentiate McGurk-failures from McGurk effects (i.e. Abbott & Shahin, 2018). Thus, EEG can provide an estimate of the timing of both the recovery of the audio-visual integrated percept and the semantic processing.

Reinterpreting Other Crossmodal Priming Findings

It is also worth reviewing how the current results fit with other studies of crossmodal speech priming. Kim, Davis, and Krins (2004) tested *repetition* priming (for which the prime and target are the same word) and found that visual-only words facilitated the identification of (the same) auditory words (see also Buchwald, Winters, & Pisoni, 2009). Interestingly however, these authors found no repetition priming effect when *nonwords* were used as the prime and target, suggesting that this crossmodal priming effect involved some lexical processing. Fort et al., (2013) found that visual-only syllables facilitated the identification of auditory-only words that started with the same syllable. Interestingly, this effect was modulated by the auditory-word's lexical frequency. While neither of these studies tested semantic priming per se, the fact that lexical characteristics (lexicality; frequency) interacted with the results suggests that

these findings are consistent with the present study in showing that visual speech affects the lexical processing of auditory speech[3].

However, in another condition, Kim et al., (2004) failed to find semantic priming between visual-only primes and auditory-only targets (e.g. visual 'back' followed by auditory 'front'). Potentially, this finding not only contrasts with the results of the present investigation, but also with the results of Ostrand et al., (2016) Experiment 1 in which visual speech did support semantic priming in the context of auditory-nonword + visual-word primes. However, Kim et al., (2004) may have failed to find a semantic priming effect of visual on auditory speech because the visual-only full word identification in their study was quite low (18% correct, on average; see p. B41). Thus, while subjects could recover enough segment information from the visual words to support cross-modal repetition priming, they could not recover enough of the full words to support semantic priming. If so, then semantic priming in the Kim et al., (2004) study was simply limited by participants struggling to identify/perceive the full visual-only words, not by a failure of visual information to access the semantic process, per se.

By using McGurk-word primes, Ostrand et al's., (2016) and the present investigations circumvented this limitation of Kim et al's (2004) design, as audio-visual words are much easier to fully identify than visual-only words. That semantic priming was more consistent with the visual, than auditory, component of the McGurk words indicates that while visual speech may often require the support of auditory speech to be

[3]It may be worth noting that there was a substantial difference in the lexical frequency of our auditory words relative to the frequency of our visual words, with the auditory words being more common. The relationship between lexical frequency and the McGurk effect are explored in more detail in Chapter 3 of this dissertation.

consistently identified, the semantic process is not preferentially sensitive to auditory information.

Another point worth considering is whether the results of the current investigation have implications for Experiment 1of Ostrand et al's report (2016). In contrast to their second experiment, Experiment 1 used auditory word + visual *non*word (e.g. auditory 'beef' + visual 'deef') and auditory *non*word + visual word (e.g. auditory 'bamp' + visual 'damp') McGurk priming stimuli. This experiment found that when the audio component was a word, there were no priming differences between visual word and visual nonword conditions. In contrast, when the auditory component was a nonword, there was an effect of visual words relative to visual nonwords. These effects were interpreted as demonstrating that when the auditory stimulus was a real word, it easily initiated lexical processing, *regardless* of whether the visual stimulus was a word or nonword. However, when the visual component was the real word, it could *only* induce priming when the auditory component was a *nonword*, thereby allowing the visual word component to have extra time to enter into lexical processing.

This interpretation conflicts with the present results which show that even in auditory word contexts, semantic processing *can* be consistent with the visual word information. Perhaps a more parsimonious interpretation of Ostrand et al's (2016) Experiment 1 is that, as in their Experiment 2, these effects reflected a preponderance of McGurk perceptions consistent with the auditory component. That is, auditory-word +visual-nonword (e.g. auditory 'beef' + visual 'deef'), were perceived as the auditory words and produced priming consistent with those auditory words (priming for 'meat').

This interpretation is consistent with the results of Brancazio (2004) who reports that auditory-word + visual-nonword combinations produce higher McGurk-auditory rates. Relatedly, auditory-nonword + visual-word conditions (e.g. auditory 'bamp' + visual 'damp') likely primed for targets related to the visual word *because* those primes were likely to produce perceptions of the McGurk stimuli consistent with the visual component (i.e. perceived as 'damp'; see Brancazio 2004).

It should also be noted however, that even in contexts when the auditory-nonword + visual-word stimuli were perceived as the auditory nonword, they likely would continue to support priming to the visual word. This is because, by necessity, the auditory-nonwords were one phonetic feature away from the real word presented visually. As noted by Ostrand et al., (2016), nonwords that closely resemble real words may support semantic priming effects that are similar to that real word (i.e. 'bamp' will semantic priming targets related to 'damp'; e.g. Connine, Blasko, & Titone, 1993; Deacon, Dynowska, Ritter, & Grose-Fifer, 2004; see also Ostrand et al., 2016 for a discussion). Thus, whether perceived as 'damp' or 'bamp,' the results of Ostrand et al's, (2016) Experiment 1 were likely driven by the participants' identification of the audio-visual stimuli, and such an effect is consistent with the results of the present investigation.

In conclusion, it is likely that lexical processing *is* sensitive to the perception of the prime. This conclusion is based on the finding that strong McGurk stimuli produce semantic priming consistent with the predicted McGurk percept, as well as our finding that the frequency of that percept correlates with the amount of semantic priming that is

ultimately observed. This conclusion is consistent with the extant literature suggesting that multisensory integration occurs early, and can readily account for not only the results of the present investigation but also for the results of both experiments of Ostrand et al., (2016). That is, we believe we provide strong evidence that semantic priming is consistent with the dominant perception associated with a multisensory priming stimulus.

References

Abbott, N. T., & Shahin, A. J. (2018). Cross-modal phonetic encoding facilitates the McGurk illusion and phonemic restoration. *Journal of Neurophysiology*, *120*(6), 2988–3000. http://doi.org/10.1152/jn.00262.2018

Alsius, A., Paré, M., & Munhall, K. G. (2018). Forty years after hearing lips and seeing voices: The McGurk effect revisited. *Multisensory Research*, *31*(1–2), 111–144. http://doi.org/10.1163/22134808-00002565

Arnold, P., & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology*, *92*(2), 339–355. http://doi.org/10.1348/000712601162220

Baart, M., & Samuel, A. G. (2015). Turning a blind eye to the lexicon : ERPs show no cross-talk between lip-read and lexical context during speech sound processing. *Journal of Memory and Language*, *85*(July). http://doi.org/10.1016/j.jml.2015.06.00

Barutchu, A., Crewther, S. G., Kiely, P., Murphy, M. J., & Crewther, D. P. (2008). When /b/ill with /g/ill becomes /d/ill: Evidence for a lexical effect in audiovisual speech perception. *European Journal of Cognitive Psychology*, *20*(1), 1–11. http://doi.org/10.1080/09541440601125623

Bebko, J. M., Schroeder, J. H., & Weiss, J. A. (2014). The McGurk effect in children with autism and asperger syndrome. *Autism Research*, *7*(1), 50–59. http://doi.org/10.1002/aur.1343

Besle, J., Fischer, C., Lecaignard, F., Bidet-Caulet, A., Lecaignard, F., Bertrand, O., & Giard, M. (2008). Visual activation and audiovisual interactions in the auditory cortex during speech perception : Intracranial recordings in humans. *The Journal of Neuroscience*, *28*(52), 14301–14310. http://doi.org/10.1523/JNEUROSCI.2875-08.2008

Brancazio, L. (2004). Lexical influences in audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(3), 445–463. http://doi.org/10.1037/0096-1523.30.3.445

Brancazio, L., & Miller, J. L. (2005). Use of visual information in speech perception: Evidence for a visual rate effect both with and without a McGurk effect. *Perception and Psychophysics*. http://doi.org/10.3758/BF03193531

Buchwald, A. B., Winters, S. J., & Pisoni, D. B. (2009). Visual speech primes open-set recognition of spoken words. *Language and Cognitive Processes*, *24*(4), 580–610. http://doi.org/10.1080/01690960802536357

Burnham, D., & Dodd, B. (2004). Auditory-visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology*, *45*(4), 204–20. http://doi.org/10.1002/dev.20032

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., … David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science*, *276*(5312), 593–596. http://doi.org/10.1126/science.276.5312.593

Cohen, J. (1992). A Power Primer. *Psychological Bulletin*, *112*(1), 155–159.

Connine, C. M., Blasko, D. G., & Titone, D. (1993). Do the beginnings of spoken words have a special status in auditory word recognition? *Journal of Memory and Language*, *32*, 193–210. http://doi.org/10.1006/jmla.1993.1011

Deacon, D., Dynowska, A., Ritter, W., & Grose-Fifer, J. (2004). Repetition and semantic priming of nonwords: Implications for theories of N400 and word recognition. *Psychophysiology*, *41*(1), 60–74. http://doi.org/10.1111/1469-8986.00120

Delaney-Busch, N., Lau, E., Morgan, E., & Kuperberg, G. R. (2017). Comprehenders Rationally Adapt Semantic Predictions to the Statistics of the Environment: A Bayesian Model of trial-level N400 amplitudes. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, 71635.

Dias, J. W., & Rosenblum, L. D. (2011). Visual influences on interactive speech alignment. *Perception*, *40*, 1457–1466. http://doi.org/10.1068/p7071

Diedenhofen, B., & Musch, J. (2015). Cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS ONE*, *10*(4), 1–12. http://doi.org/10.1371/journal.pone.0121945

Fort, M., Kandel, S., Chipot, J., Savariaux, C., Granjon, L., & Spinelli, E. (2013). Seeing the initial articulatory gestures of a word triggers lexical access. *Language and Cognitive Processes*, *28*(8), 1–17. http://doi.org/10.1080/01690965.2012.701758

Fowler, C. A., Brown, J. M., & Mann, V. A. (2000). Contrast effects do not underlie effects of preceding liquids on stop-consonant identification by humans. *Journal of Experimental Psychology. Human Perception and Performance*, *26*(3), 877–888. http://doi.org/10.1037/0096-1523.26.3.877

Green, K. P., & Kuhl, P. K. (1989). The role of visual information in the processing of place and manner features in speech perception. *Perception & Psychophysics*, *45*(1), 34–42. http://doi.org/10.3758/BF03208030

Green, K. P., & Miller, J. L. (1985). On the role of visual rate information in phonetic

perception. *Perception & Psychophysics*, *38*(3), 269–276. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/4088819

Green, K. P., & Norrix, L. W. (2001). Perception of /r/ and /l/ in a stop cluster: Evidence of crossmodal context effects. *Journal of Speech, Language and Hearing Reseach*. *37*(1) 166-177. http://doi.org/10.1044/jslhr.4003.646

Jordan, T. R., McCotter, M. V., & Thomas, S. M. (2000). Visual and audiovisual speech perception with color and gray-scale facial images. *Perception and Psychophysics*, *62*(7), 1394–1404. http://doi.org/10.3758/BF03212141

Kim, J., Davis, C., & Krins, P. (2004). Amodal processing of visual speech as revealed by priming. *Cognition*, *93*(1). http://doi.org/10.1016/j.cognition.2003.11.003

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. *The Computer and Literary Studies* 153-165.

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics*, *24*(3), 253–7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/704285

Magnotti, J. F., & Beauchamp, M. S. (2015). The noisy encoding of disparity model of the McGurk effect. *Psychonomic Bulletin & Review*, *22*(3), 701–709. http://doi.org/10.3758/s13423-014-0722-2

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1–2), 71–102.

Milberg, W., Blumstein, S. E., & Dworetzky, B. (1988). Phonological factors in lexical access: Evidence from an auditory lexical decision task. *Bulletin of the Psychonomic Society*, *26*(4), 305–308.

Mitterer, H., & Reinisch, E. (2017). Visual speech influences speech perception immediately but not automatically. *Attention, Perception, and Psychophysics*, *79*(2), 660–678. http://doi.org/10.3758/s13414-016-1249-6

Musacchia, G., Sams, M., Nicol, T., & Kraus, N. (2006). Seeing speech affects acoustic information processing in the human brainstem. *Experimental Brain Research*, *168*(1–2), 1–10. http://doi.org/10.1007/s00221-005-0071-5

Namasivayam, A. K., Yiu, W., & Wong, S. (2015). Visual speech gestures modulates the efferent auditory system, *14*(1), 73–83. http://doi.org/10.1142/S0219635215500016

Navarra, J., & Soto-Faraco, S. (2007). Hearing lips in a second language: Visual articulatory information enables the perception of second language sounds. *Psychological Research*, *71*(1), 4–12. http://doi.org/10.1007/s00426-005-0031-5

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, *106*(3), 226–254.

Nelson, D., McEvoy, C. L., & Schreiber, T. A. (1998). The university of south florida word association, rhyme, and word fragment norms, *36*(3). Retrieved from http://www.usf.edu/FreeAssociation/

Ostrand, R., Blumstein, S. E., Ferreira, V. S., & Morgan, J. L. (2016). What you see isn't always what you get: Auditory word signals trump consciously perceived words in lexical access. *Cognition*, *151*, 96–107. http://doi.org/10.1016/j.cognition.2016.02.019

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I. P., Möttönen, R., Tarkiainen, A., & Sams, M. (2005). Primary auditory cortex activation by visual speech: An fMRI study at 3 T. *Neuroreport*, *16*(2), 125–128. http://doi.org/10.1097/00001756-200502080-00010

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A speechreading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97–113). London, England: Erlbaum.

Rosenblum, L.D. (In press). Audiovisual Speech Perception and the McGurk Effect. In *The encyclopedia of applied linguistics*.

Rosenblum, L. D., Dias, J. W., & Dorsi, J. (2016). The supramodal brain: Implications for auditory perception. *Journal of Cognitive Psychology*, *5911*, 1–23. http://doi.org/10.1080/20445911.2016.1181691

Rosenblum, L. D., & Saldaña, H. M. (1992). Discrimination tests of visually influenced syllables. *Perception & Psychophysics*, *52*(4), 461–473. http://doi.org/10.3758/BF03206706

Saldaña, H. M., & Rosenblum, L. D. (1994). Selective adaptation in speech perception using a compelling audiovisual adaptor. *The Journal of the Acoustical Society of America*, *95*(6), 3658–3661. http://doi.org/10.1121/1.409935

Sams, M., Manninen, P., Surakka, V., Helin, P., & Kättö, R. (1998). McGurk effect in Finnish syllables, isolated words, and words in sentences: Effects of word meaning

and sentence context. *Speech Communication*, *26*(1–2), 75–87. http://doi.org/10.1016/S0167-6393(98)00051-X

Samuel, A. G., & Lieblich, J. (2014). Visual speech acts differently than lexical context in supporting speech perception. *Journal of Experimental Psychology. Human Perception and Performance*, *40*(4), 1479–90. http://doi.org/10.1037/a0036656

Sanchez, K., Miller, R. M., & Rosenblum, L. D. (2010). Visual influences on alignment. *Journal of Speech, Language, and Hearing Research*, *53*, 262–272.

Shahin, A. J., Backer, K. C., Rosenblum, L. D., & Kerlin, J. R. (2018). Neural mechanisms underlying cross-modal phonetic encoding. *The Journal of Neuroscience*, *38*(7), 1566–17. http://doi.org/10.1523/JNEUROSCI.1566-17.2017

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*(2), 212–215.

Strand, J., Cooperman, A., Rowe, J., & Simenstad, A. (2014). Individual differences in susceptibility to the McGurk effect: Links with lipreading and detecting audiovisual incongruity. *Journal of Speech, Language and Hearing Reseach*, *57*, 2322–2331. http://doi.org/10.1044/2014

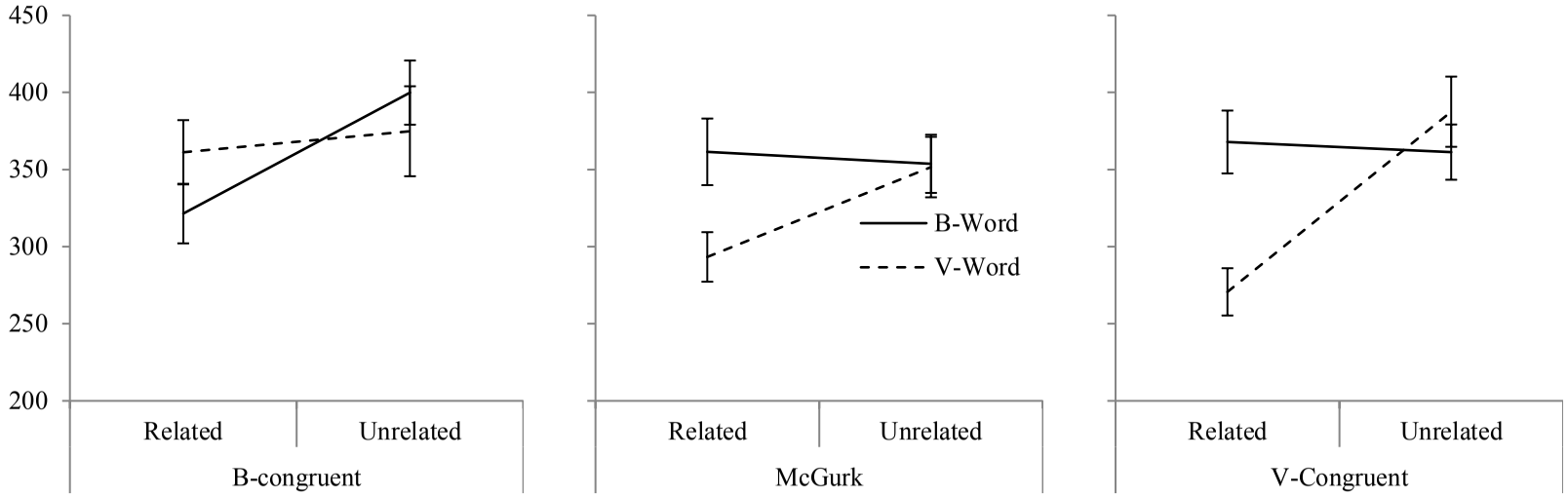
Stropahl, M., Schellhardt, S., & Debener, S. (2016). McGurk stimuli for the investigation of multisensory integration in cochlear implant users: The Oldenburg Audio Visual Speech Stimuli (OLAVS). *Psychonomic Bulletin & Review*, 1–10. http://doi.org/10.3758/s13423-016-1148-9

Teinonen, T., Aslin, R. N., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition*, *108*(3), 850–855. http://doi.org/10.1016/j.cognition.2008.05.009

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, *45*(3), 598–607. http://doi.org/10.1016/j.neuropsychologia.2006.01.001

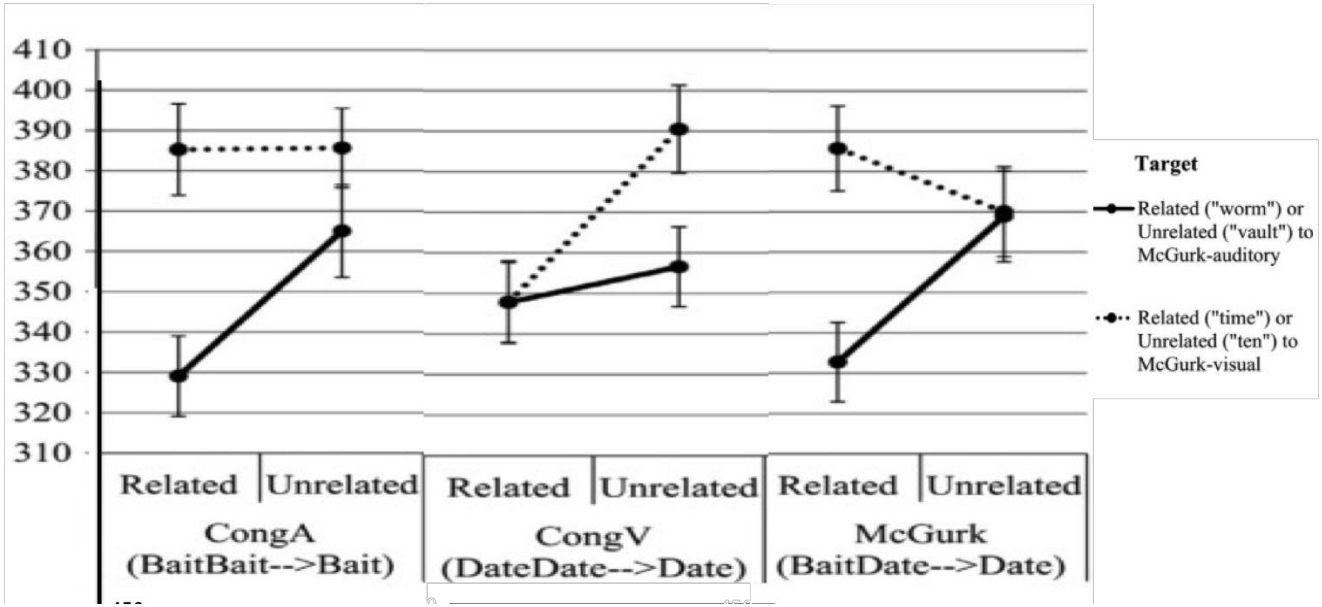
Windmann, S. (2004). Effects of sentence context and expectation on the McGurk illusion. *Journal of Memory and Language*, *50*(2), 212–230. http://doi.org/10.1016/j.jml.2003.10.001

Figure 1.1a



The values on the vertical axis are reaction times following target offset. Solid lines correspond to targets related or unrelated to the McGurk auditory word. Broken lines correspond to targets related or unrelated to the McGurk visual word. Error bars show the standard error of the mean. Data were tabulated by subjects (i.e. F_1 Analysis). The slope of each line indicates the reaction time difference between targets related and unrelated to the preceding prime. Steeper slopes indicate larger differences, and postive slopes indicate shorter reaction times to related than to unrelated targets. Column 1 displays the data for audio-visually congruent primes consistent with the McGurk auditory words. The flat slope of the broken line indicates that these stimuli do not semantically prime for targets associated with the McGurk visual words. The slope of the solid line indicates that there is semantic priming for targets associated with the McGurk auditory word primes. Column 2 displays data for audio-visually congruent primes consistent with the McGurk visual words. The positive slope of the broken line indicates there is semantic priming for targets associated with the McGurk visual prime words. The flat slope of the solid line indicates that there is no semantic priming for targets associated with the McGurk auditory word primes. Column 3 shows McGurk (audio-visually incongruent) primes. The slope of the broken line indicates semantic priming for the visual stimulus, while the solid line indicates a lack of priming for the auditory stimulus.

Figure 1.1b



Taken from Ostrand et al., (2016), Figure 1.1b can be interpreted in the same way as Figure 1.1a. Note that the first two coloumns show a similar pattern between figures; Column 1 shows a flat broken line and a postively sloped solid line (corresponding to auditory and visual associated targets respectively), Column 2 shows a positively sloped broken line and a flat solid line. The key difference between these figures is in Column 3. In Figure 1.1a Column 3 approximates the relationship shown in Column 2 suggesting that McGurk stimuli produce priming similar to primes that are audio-visually congruent with their visual stimuli. In contrast, in Figure 1.1b Column 3 approximates the pattern of Column 1, indicating that that McGurk stimuli produce priming similar to primes that are audio-visually congruent with their auditory stimuli.

Figure 1.2

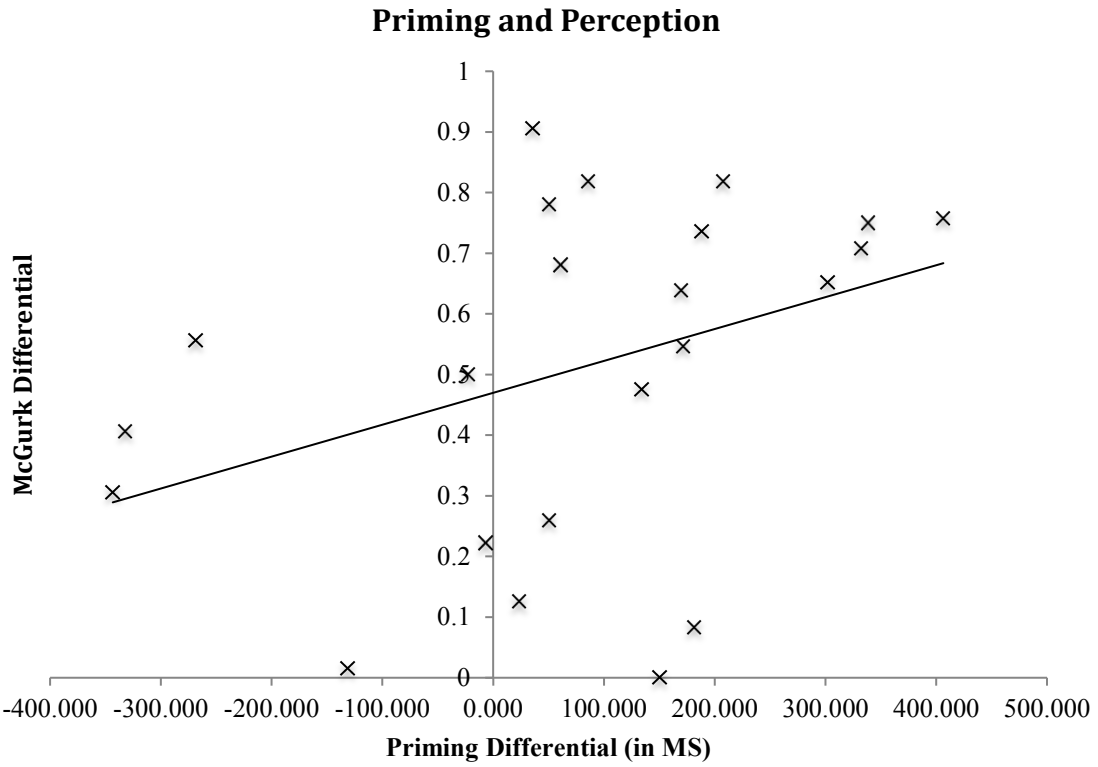
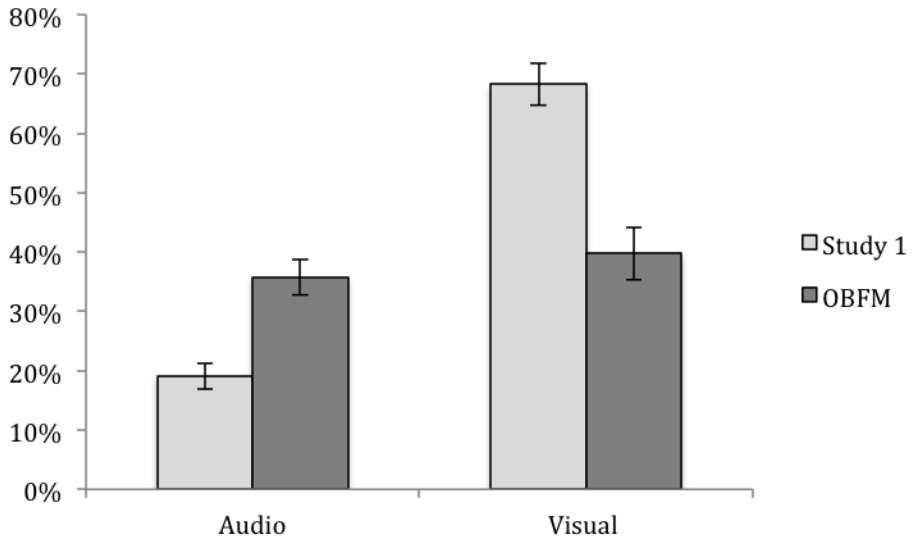


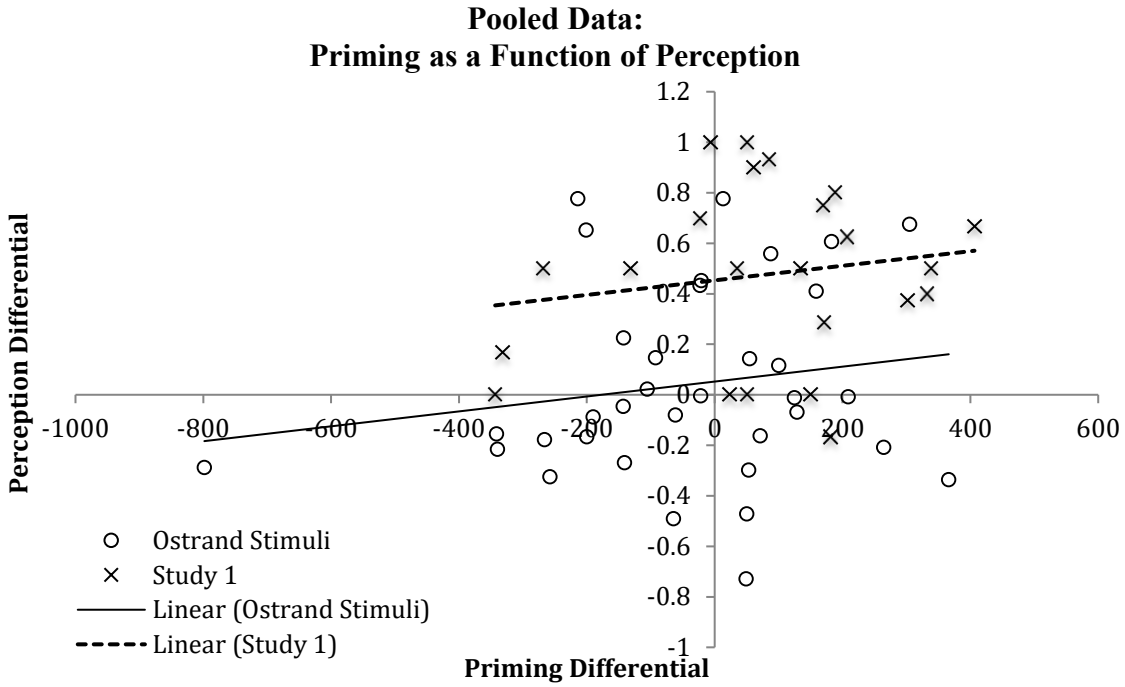
Figure 1.2 shows the relationship between semantic priming and the McGurk effect. The vertical axis shows the rate of McGurk-visual responses minus the rate of McGurk-auditory responses for each item. Negative numbers indicate that the item produced more McGurk-auditory (perceptions of the auditory word) than McGurk-visual responses. The McGurk values exclude responses from participants who did not contribute reaction times to the priming calculation. The priming differential is shown along the horizontal axis and was calculated by subtracting the difference of reaction times between targets unrelated and related to the McGurk auditory word from the difference of reaction times between targets unrelated and related to the McGurk visual word. Negative numbers indicate that the priming effect of the McGurk auditory words was larger than the priming effect of the McGurk visual word.

Figure 1.3



This figure shows the rates of McGurk-visual and McGurk-auditory (perception of the auditory channel) responses for the stimuli used by Ostrand, Blumstein, Ferreira, and Morgan (2016; 'OBFM'; dark bars) and Study 1 (light bars). McGurk rates are shown in percentage of initial consonant responses that were consistent with either the auditory (left panel) or visual (right panel) words for the McGurk stimuli averaged across items. Error bars refer to the standard error of the mean.

Figure 1.4



This figure shows the relationship between semantic priming and McGurk perception for both Study 1 and Ostrand et al., (2016). The vertical axis shows the McGurk differential, again calculated as the difference between McGurk-visual and McGurk-auditory responses. In contrast to Figure 1.2, the McGurk differential used in Figure 1.4 only includes perception responses from participants *who did not* contribute reaction times to the priming calcuation. Negative numbers indicate more frequent McGurk-auditory than McGurk-visual responses. The horizontal axis shows the priming differential, calculated by subtracting difference of reaction times between targets unrelated and related to the McGurk auditory word from the difference of reaction times between targets unrelated and related to the McGurk visual word. Negative numbers indicate that the priming effect of the McGurk auditory words was larger than the priming effect of the McGurk visual word. Crosses show the items from Study 1 and circles show the items from Ostrand et al., (2016). The solid trend line shows the trend for Study 1 while the broken line refers to Ostrand et al., (2016).

Table 1.1

Prime		Audio Associates		Visual Associates	
Audio	Visual	Related	Unrelated	Related	Unrelated
Bale	Veil	Hay	Song	Wedding	Want
Ban	Van	Stop	Sell	Car	True
Bane	Vein	Curse	Dance	Blood	Parking
Base	Vase	Bottom	Fender	Flowers	Seller
Bat	Vat	Ball	Movement	Tub	Strength
Beer	Veer	Drink	Smaller	Swerve	Letter
Bending	Vending	Break	Dig	Machine	Much
Bent	Vent	Broken	Exile	Air	Disappear
Best	Vest	Worst	Stop	Clothes	Car
Bet	Vet	Money	Hay	Animals	Wedding
Bile	Vial	Stomach	Curse	Potion	Blood
Boat	Vote	Water	Bottom	Elect	Flowers
Bolt	Volt	Nut	Ball	Shock	Tub
Bow	Vow	Down	Drink	Marriage	Swerve
Bowl	Vole	Dish	Break	Mouse	Machine
Burst	Versed	Explode	Broken	Well	Clothes
Buy	Vie	Sell	Break	Want	Air
Ballad	Valid	Song	Money	True	Animals
Ballet	Valet	Dance	Stomach	Parking	Potion
Bender	Vendor	Fender	Water	Seller	Elect
Bigger	Vigor	Smaller	Nut	Strength	Shock
Bowel	Vowel	Movement	Down	Letter	Marriage
Bury	Very	Dig	Dish	Much	Well
Banish	Vanish	Exile	Explode	Disappear	Mouse

Column 1 of Table 1.1 shows the words used in the b-congurent prime conditions, and that are used for the McGurk auditory stimulus. Column 2 shows the words used in the v-congruent prime conditions and that are used for the McGurk visual stimulus. The McGurk stimulus combines words from columns 1 and 2 within each row. The remaining columns of each row display the target words that are related and unrelated to the words in columns 1 and 2.

Table 1.2

By Subjects

Effect	McGurk vs. CongA	McGurk vs. CongV	CongA vs. CongV
Related	$F(1, 98)= 9.746$, $p = .002$, $\eta^2_p = .090$*	$F(1, 98)= 14.449$, $p < .001$, $\eta^2_p = .128$*	$F(1, 95)= 19.740$, $p < .001$, $\eta^2_p = .172$*
Target	$F(1, 98)= 5.203$, $p = .025$, $\eta^2_p = .050$*	$F(1, 98)= 16.535$, $p < .001$, $\eta^2_p = .144$*	$F(1, 95)= 4.311$, $p = .041$, $\eta^2_p = .043$*
Prime	$F(1, 98)= 2.719$, $p = .102$, $\eta^2_p = .027$	$F(1, 98)= .247$, $p = .621$, $\eta^2_p = .003$	$F(1, 95)= 1.436$, $p = .234$, $\eta^2_p = .015$
Related x Target	$F(1, 98)=.065$, $p = .800$, $\eta^2_p = .001$	$F(1, 98)= 14.828$, $p < .001$, $\eta^2_p = .131$*	$F(1, 95)= .985$, $p = .323$, $\eta^2_p = .010$
Related x Prime	$F(1, 98)= 1.223$, $p = .272$, $\eta^2_p = .012$	$F(1, 98)= .897$, $p = .346$, $\eta^2_p = .009$	$F(1, 95)= 0.013$, $p = .910$, $\eta^2_p < .001$
Target x Prime	$F(1, 98)= .927$, $p = .338$, $\eta^2_p = .009$	$F(1, 98)< .001$, $p = .984$, $\eta^2_p < .001$	$F(1, 95)= 2.381$, $p = .126$, $\eta^2_p = .024$
3 way	**$F(1, 98)= 11.166$, $p = .001$, $\eta^2_p = .102$***	$F(1, 98)= 1.306$, $p = .256$, $\eta^2_p = .013$	**$F(1, 95)= 17.044$, $p < .001$, $\eta^2_p = .152$***

By Items

Effect	McGurk vs. CongA	McGurk vs. CongV	CongA vs. CongV
Related	$F(1, 23)= 2.658$, $p = .117$, $\eta^2_p = .104$	$F(1, 23)= 5.074$, $p = .034$, $\eta^2_p = .181$*	$F(1, 23)= 5.473$, $p = .028$, $\eta^2_p = .192$*
Target	$F(1, 23)=.521$, $p = .478$, $\eta^2_p = .022$	$F(1, 23)= 3.418$, $p = .077$, $\eta^2_p = .129$	$F(1, 23)= 1.270$, $p = .271$, $\eta^2_p = .052$
Prime	$F(1, 23)= 7.090$, $p = .014$, $\eta^2_p = .236$*	$F(1, 23)=.603$, $p = .445$, $\eta^2_p = .026$	$F(1, 23)= 2.073$, $p = .163$, $\eta^2_p = .083$
Related x Target	$F(1, 23)=.060$, $p = .809$, $\eta^2_p = .003$	$F(1, 23)= 18.793$, $p < .001$, $\eta^2_p = .450$*	$F(1, 23)= 2.431$, $p = .133$, $\eta^2_p = .096$
Related x Prime	$F(1, 23)= .667$, $p = .423$, $\eta^2_p = .028$	$F(1, 23)= 2.011$, $p = .170$, $\eta^2_p = .080$	$F(1, 23)=.191$, $p = .667$, $\eta^2_p = .008$
Target x Prime	$F(1, 23)= 1.168$, $p = .291$, $\eta^2_p = .048$	$F(1, 23)=.429$, $p = .519$, $\eta^2_p = .018$	$F(1, 23)= 3.461$, $p = .076$, $\eta^2_p = .131$
3 way	**$F(1, 23)= 7.402$, $p = .012$, $\eta^2_p = .243$***	$F(1, 23)= 1.615$, $p = .216$, $\eta^2_p = .066$	**$F(1, 23)= 22.905$, $p < .001$, $\eta^2_p = .499$***

The top panel to Table 1.2 shows the results from the F_1 analyses examining two levels of the prime condition, the bottom panel displays the results from the F_2 analyses. In both panels, the first column shows results when the prime factor included McGurk primes and words that were audio-visually congruent to the McGurk auditory channel (e.g. Column 1 of Table 1.1). The second column shows the results when the prime factor included the McGurk primes and words that were audio-visually congruent with the McGurk visual channel (e.g. Column 2 of Table 1.1). The third column shows the results when prime included the two audio-visually congruent conditions. Bolded results were significant. The critical result is the 3-way interaction which is present for the McGurk vs. b-congruent and b-congruent vs. v-congruent but not the McGurk vs. v-congruent columns for both panels.

Table 1.3

Auditory Word	McGurk % Visual ('V')	McGurk % Auditory ('B')	Congruent (% 'V') B words	Congruent (% 'V') V words	Audio-Only (% 'V') B words	Audio-Only (% 'V') V words
Bane	78	13	25	91	45	81
Bowel	86	12	19	98	30	94
Bile	76	9	20	98	33	98
Bet	82	14	12	96	22	96
Ban	86	7	7	95	4	96
Bender	64	29	11	89	9	87
Burst	26	24	6	83	13	85
Bale	79	4	30	95	68	94
Bigger	59	30	17	82	12	83
Base	75	8	10	95	55	94
Beer	19	10	1	16	2	17
Bowl	56	29	8	90	11	86
Ballad	77	20	8	99	22	98
Boat	65	32	11	98	42	98
Bury	49	48	14	84	14	88
Banish	71	26	15	97	28	98
Best	80	14	2	96	26	90
Ballet	69	29	10	96	14	96
Bent	79	17	7	98	32	97
Bat	65	8	13	76	36	59
Bow	76	22	5	89	11	76
Bolt	80	15	27	98	35	98
Bending	76	18	6	98	14	97

Column 1 of Table 1.3 shows the auditory words of the McGurk stimuli (corresponding visual words are shown in column 2 of Table 1.1). Column 2 shows the percentage of 'V' initial responses to each McGurk item, and column 3 shows the percentage of 'B' initial responses. The remaining columns show the percentage of 'V' initial responses for audio-visual congruent and audio-alone, B words and V words.

Table 1.4

Effects	Results from Omnibus ANCOVA
Target	$F(1, 21)= 1.024, p = .323, \eta^2_p = .046$
*Target x McGurk	$F(1, 21)= 4.539, p = .045, \eta^2_p = .178$
Related	$F(1, 21)= 1.094, p = .307, \eta^2_p = .050$
Related x McGurk	$F(1, 21)=.002, p = .966, \eta^2_p = .000$
*Prime	$F(2, 42)= 5.086, p = .011, \eta^2_p = .195$
*Prime x McGurk	$F(2, 42)= 4.410, p = .018, \eta^2_p = .174$
Target x Related	$F(1, 21)= 1.023, p = .323, \eta^2_p = .046$
Target x Related x McGurk	$F(1, 21)= .006, p = .939, \eta^2_p = .000$
Target x Prime	$F(2, 42)= .759, p = .474, \eta^2_p = .035$
Target x Prime x McGurk	$F(2, 42)= .102, p = .904, \eta^2_p = .005$
Related x Prime	$F(2, 42)= .070, p = .932, \eta^2_p = .003$
Related x Prime x McGurk	$F(2, 42)= .425, p = .657, \eta^2_p = .020$
*Target x Prime x Related	$F(2, 42)= 4.687, p = .015, \eta^2_p = .182$
***Target x Prime x Related x McGurk**	$\mathbf{F(2, 42)= 3.614, p = .036, \eta^2_p = .147}$

Table 1.4 displays the results of the Target (V word vs. B word associates) x Related (related vs. unrelated) x Prime (McGurk vs. CongV vs. CongA) x McGurk rate ACOVA. The key 4-way interaction is shown in bolded font. Other significant effects are indicated by astrisks.

Table 1.5

	Ostrand et al., (2016)		Study 1				Comparison
	F-value	*r*-value	*F*-value	*r*-value	CI - High (*r*)	CI - Low (*r*)	Ostrand et al., (2016) in Study 1 CI?
Related	13.44	0.298	19.74	0.415	0.565	0.226	Yes
Target	**25.34**	**0.394**	**4.311**	**0.208**	**0.389**	**0.000**	**No**
Prime	1.327	0.098	1.436	0.122	0.310	-0.080	Yes
Related x Target	0.005	0.006	0.985	0.101	0.291	-0.100	Yes
Related X Prime	0.545	0.063	0.013	0.012	0.207	-0.187	Yes
Target x Prime	3.834	0.164	2.381	0.156	0.336	-0.050	Yes
3-way	12.069	0.284	17.044	0.390	0.551	0.017	Yes

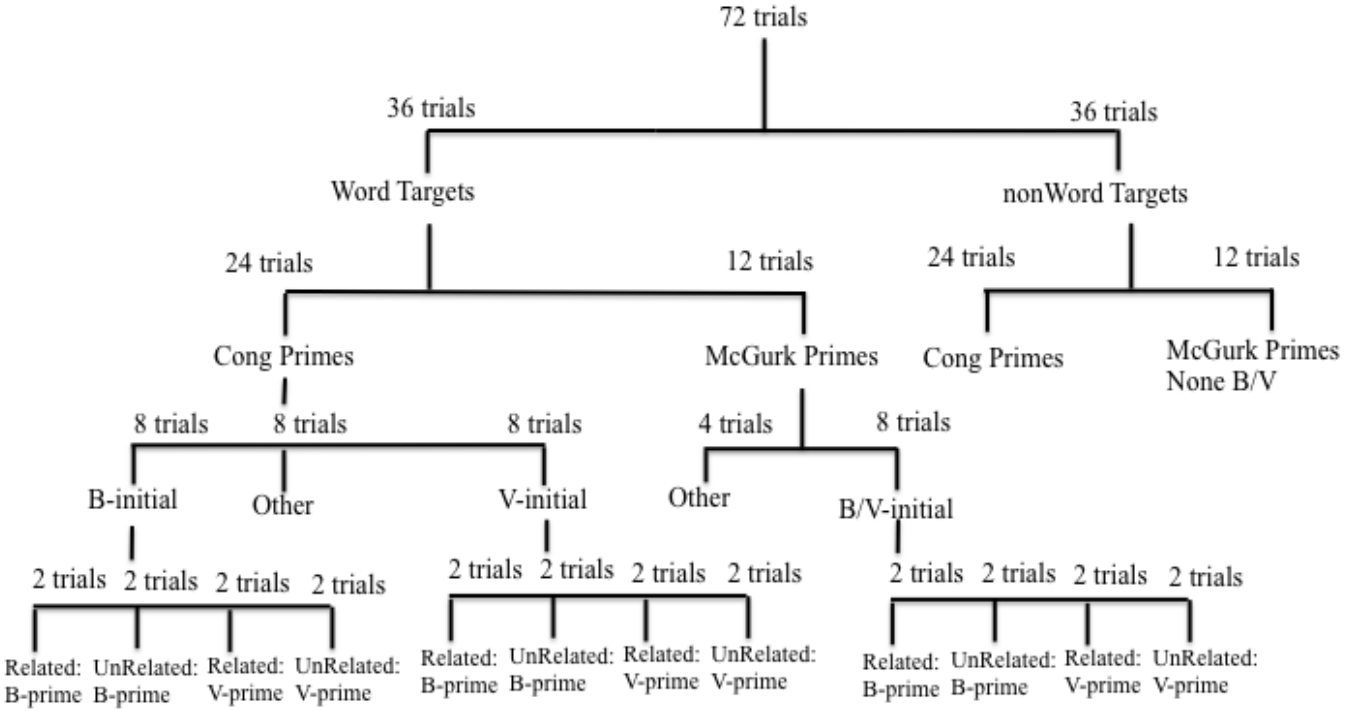
This table displays the *F*-values reported by Ostrand et al., (2016) and the corresponding effect sizes (calculated as *r*) in columns 2 and 3. Columns 4-7 show the corresponding values from Study 1 as well as the 95% confidence interval around Study 1's effect size. Column 8 indicates which results of Ostrand et al., (2016) were within the confidence interval of Study 1. Note that only the effect of target is outside Study 1's confidence interval and is only marginally (.005) so.

Table 1.6

Stimuli		Perception	
Visual Word	Auditory Word	% Auditory	% Visual
Dad	Bad	26	71
Date	Bait	30	53
Dank	Bank	72	25
Day	Bay	22	5
Deed	Bead	18	10
Dean	Bean	41	56
Deer	Beer	30	14
Dell	Bell	57	41
Debt	Bet	42	13
Did	Bid	49	28
Dill	Bill	61	12
Tad	Pad	21	20
Tart	Part	25	25
Tie	Pie	45	47
Toll	Pole	28	22
Tot	Pot	45	44
Tug	Pug	41	55
Test	Pest	57	41
Nail	Mail	22	77
Name	Maim	19	80
Nap	Map	16	83
Knee	Me	10	88
Nice	Mice	10	88
Night	Might	52	47
Nil	Mill	15	81
Nine	Mine	18	61
Nix	Mix	28	69
Nob	Mob	42	42
Node	Mode	38	17
Crimp	Primp	86	13
Gore	Bore	18	10
Gum	Bum	53	23
Gun	Bun	44	12
Cod	Pod	47	13
Gut	Butt	48	22
Guy	Buy	13	25

Columns 1 and 2 of Table 1.6 show the auditory and visual words (respectively) of the McGurk stimuli used by Ostrand et al., (2016). Columns 3 and 4 show the corresponding auditory and visual percpeitons for these stimuli, as indicated by percent of reported initial consonant consistent with the auditory or visual stimulus.

Appendix A


72 trials
36 trials
36 trials
Word Targets
nonWord Targets
24 trials
12 trials
24 trials
12 trials
Cong Primes
McGurk Primes
Cong Primes
McGurk Primes
None B/V
8 trials
8 trials
8 trials
4 trials
8 trials
B-initial
Other
V-initial
Other
B/V-initial
2 trials 2 trials 2 trials 2 trials
2 trials 2 trials 2 trials 2 trials
2 trials 2 trials 2 trials 2 trials
Related: B-prime
UnRelated: B-prime
Related: V-prime
UnRelated: V-prime
Related: B-prime
UnRelated: B-prime
Related: V-prime
UnRelated: V-prime
Related: B-prime
UnRelated: B-prime
Related: V-prime
UnRelated: V-prime

Chapter 2

Cross-Modal Information for Phonemic Restoration:

Insights Offered by Selective Adaptation

Cross-Modal Information for Phonemic Restoration: Insights Offered by Selective Adaptation

In most circumstances, individuals must perceive speech against a variety of environmental noise including sounds from office work, nearby traffic, and other talkers. Accurate speech perception in the complex and dynamic environment in which speech perception takes place is often aided by contextual information that accompanies the speech auditory signal. This includes multisensory information, such as the visible articulations that accompany the auditory signal, as well as lexical information provided by the words in which each audible segment occurs.

Speech is somewhat unique in that it is both an event that occurs in the environment and also a message sent between interlocutors. That is, speech is processed both perceptually, to determine what articulatory event occurred, and linguistically to determine what meaning was conveyed by that event. That both multisensory and lexical (word context) information support speech perception is well illustrated by speech in noise listening tasks; listeners are more accurate at identifying segments when they can see the talker (Sumby & Pollack, 1954; Grant & Seitz, 2000) or when that talker is saying words as opposed to nonwords (Miller, Heise, & Lichten, 1951; Hirsh, Reynolds, & Joseph, 1954). But do these influences on speech identification necessarily imply that both lexical and multisensory information influence *linguistic* processing as well?

Samuel and Lieblich (2014) argue that the answer to this question is no. They base their argument on a pattern of findings concerning a phenomenon known as selective adaptation (Emias & Corbit, 1973). Selective adaptation is a finding that

repeated exposure to a speech stimulus will change subsequent speech perception, such that fewer speech stimuli will be identified as belonging to the phonetic category of the previously presented item (Eimas & Corbit, 1973). For example, following 150 rapidly presented /pa/ tokens, fewer items from a /ba/-/pa/ continuum will be identified as /pa/ (Eimas & Corbit, 1973). Based on this characteristic result, selective adaptation is sometimes described as a fatiguing of phonetic detectors or retuning phoneme classification criteria (e.g. Samuel, 1986; Kleinschmidt & Jaeger, 2016). For Samuel and Lieblich (2014), selective adaptation reflects true linguistic processing, whereas simple phonetic identifications may be influenced by decision bias. Importantly, Samuel and Lieblich (2014) claim that the literature shows a dissociation in selective adaptation effects for speech in lexical contexts relative to multisensory contexts.

Samuel (1997) measured selective adaptation resulting from a lexically supported speech illusion, the phonemic restoration effect. Phonemic restoration was first reported by Warren (1970) who removed a segment from a word utterance and replaced this segment with noise (e.g. Warren replaced the central 's' of "legi**s**latures" with a coughing sound). Warren (1970) found that in these conditions listeners erroneously reported hearing the speech segment that had been removed. Samuel (1997) used phonemic restoration stimuli as the repetitively presented items in a selective adaptation paradigm. These phonemic restoration stimuli were words with either a central /d/ or /b/ segment replaced by noise (e.g. 'arma#ilo' & 'inhi#ition'). Accordingly, the test continuum on which Samuel measured adaptation was a /bi/-/di/ continuum. Samuel found that presenting noise-replace /b/ words resulted in fewer items on the /bi/-/di/ continuum

being identified as /bi/ (/bi/ adaptation), and presenting noise-replace /d/ words resulted in fewer items on the /bi/-/di/ continuum being identified as /di/ (/di/ adaptation). In finding selective adaptation, Samuel concluded that the lexical context served to phonemically restore the missing segments, and supported true linguistic processing.

Following this report, Samuel (2001) further investigated the lexical sensitivity of selective adaptation. In this study, a single speech segment that was ambiguous between /s/ and /ʃ/ (henceforth /?/) was appended to audible /s/ and /ʃ/ biasing word segments; such as "Christma" (as in "Christma**s**") or "Demoli" (as in "Demoli**sh**"). These stimuli produced the classic Ganong effect (Ganong, 1980), the word context in which /?/ was inserted determined how /?/ was identified. More importantly, despite using the same /?/ segment in both conditions, Samuel (2001) found /s/ adaptation for "Christma/?/" and /ʃ/ for "Demoli/?/" stimuli. Interestingly, Samuel and Frost (2015) found this Ganong adaptation effect with high English proficient bilingual subjects but not with low English proficient bilinguals, thereby localizing the effects to the strength of lexical representations. Thus, across three studies, selective adaptation appears to be sensitive to lexical information.

These findings contrast with what has been found for multisensory contexts, for which McGurk stimuli have consistently failed to produce selective adaptation effects. This literature begins with a study by Roberts and Summerfield (1981) who compared the selective adaptation effects produced by auditory-only /ba/ and /da/ to audio-visually incongruent (McGurk type) adaptors. The incongruent adaptors were composed of auditory /ba/ and visual /ga/ articulations, which generally create a /da/ percept

(MacDonald & McGurk, 1978; McGurk & MacDonald, 1976). Roberts and Summerfield (1981) found strong adaptation effects for the auditory-only /ba/ and /da/ segments. Critically, while the incongruent auditory /ba/ + visual /ga/ adaptors were frequently perceived as /da/, these authors found that the McGurk adaptors produced an adaptation effect in the same direction as the audio-only /ba/ adaptor. These results indicated that subjects had only adapted to the auditory token of the incongruent stimulus, with no apparent influence of the visual context or the illusory percept.

Concerned that the findings of Roberts and Summerfield (1981) might reflect weak audio-visual integration, Saldana and Rosenblum (1994) conducted a follow-up experiment using a more compelling McGurk stimulus (auditory /ba/ & visual /va/, which was perceived as /va/ 99% of the time). However, despite these improved stimuli, Saldana and Rosenblum (1994) replicated the original finding: adaptation appeared to be driven by the unperceived auditory stimulus. These researchers concluded that poor cross-modal integration was unlikely to account for the results of Roberts and Summerfield (1981). Other studies have also found that that McGurk adaptors produce adaptation to the putatively unperceived auditory stimulus (Shigeno, 2002; van Linden, 2007; see also Samuel & Lieblich, 2014; Luttke et al., 2016).

The success of lexical context, and the failure of multisensory context, to support selective adaptation led Samuel and Lieblich (2014) to propose their account of speech processing which postulates separate perceptual and linguistic speech processes. For these authors, the perceptual process corresponds to the phenomenological experience of a speech stimulus, while the linguistic process analyzes the meaning of that stimulus.

Thus, while an individual may perceive one thing, their mind may be linguistically processing something entirely different.

As an illustrative example of their account, Samuel and Lieblich point to work by Ostrand et al., (2011; 2016) in which audio-visual words were used in a semantic priming paradigm. The audio-visual speech of Ostrand et al's (2016) study included McGurk words, in which the auditory stimulus and the perception of that stimulus could be two different words. The essential finding of this research was that semantic processing was consistent with the auditory, as opposed to the visual (and putatively perceived) signal in McGurk stimuli (but see Chapter 1). These results are consistent with the two parallel processes proposed by Samuel and Lieblich: the perceptual process produced the phenomenological experience of the McGurk words—the participants perceived the visual stimulus—while the linguistic process, concerned the *meaning* of the unperceived auditory component of the McGurk words[4].

For Samuel and Lieblich (2014) selective adaptation operates similarly to semantic priming, being sensitive to linguistic, rather than perceptual, processing. While semantic priming clearly reflects processing of linguistic meaning, selective adaptation is sensitive to low-level phonetic information. For example, selective adaptation is typically found in response to isolated syllables (Eimas & Corbit, 1973). Moreover, selective adaptation will generalize across phonetic categories on the basis of shared feature

[4] While Samuel and Lieblich (2014) explicitly state that "the speech signal is both a perceptual object, and a linguistic object." (p. 1488) and that these two "objects" interact with corresponding speech processes: "…functionally separate processing of the linguistic and perceptual aspects of spoken language …" (p. 1489), they do not provide a clear explanation of the distinction between a "perceptual object" and a "linguistic object." Based on a careful reading of their argument, we feel the "phenomenological" vs. "meaning" distinction accurately captures their intentions.

information (Eimas & Corbit, 1973). Selective adaptation can even be driven by non-speech stimuli that approximate speech features (i.e. noise will produce fricative adaptation; Samuel & Newport, 1979). Collectively, these findings suggest that selective adaptation is sensitive to *early stages of speech processing*: earlier than phoneme recovery. That Samuel and Lieblich (2014) argue that selective adaptation is diagnostic of a separation of perceptual and linguistic processes indicates that this separation likewise also occurs *early*.

This proposal, that linguistic processes are insensitive to multisensory information, is surprising for a number of reasons. First, if multisensory context can influence the recovery of linguistic information, as it does in the McGurk effect, then why would linguistic processing operate independently of multisensory perception? Second, there is evidence that visual speech *can* crossmodally influence linguistic processing (Fort et al., 2013; Kim, Davis, & Krins, 2004). Specifically, Kim et al., (2004) found repetition priming for lip-read utterances preceding auditory-only utterances. Importantly, this effect was only found for repeated words, not nonwords. Thus, this cross-sensory facilitation involved linguistic processing. Similarly, Fort et al., (2013) found that lip-read syllables facilitated the identification of auditory words that shared the same initial segment (phonological priming) but the amount of this facilitation was mediated by the lexical frequency of the auditory word. Like Kim et al., (2004) this work shows that linguistic processing is responsive to cross-sensory information. Moreover, as reported in Chapter 1 of this dissertation, multisensory information can influence semantic priming, a paradigm that Samuel and Lieblich (2014) argue reflects linguistic

processing and that should be insensitive to multisensory information. These findings all appear contrary to the account offered by Samuel and Lieblich (2014).

Finally, the Samuel and Lieblich (2014) account is surprising because there is now ample evidence that multisensory integration occurs early, likely before other steps in the language process (See Rosenblum et al., 2016 for a review). To briefly summarize, there is evidence of visual speech activating the primary auditory cortex (Calvert et al., 1997; Okada, Venezia, Matchin, Saberi, & Hickok, 2013; Pekkola et al., 2005; Besle et al., 2004; and for a review, see Rosenblum et al., 2016) as well of modulations of auditory processing in multisensory contexts (Stekelenburg & Vroomen, 2012; 2007; Vroomen & Stekelenburg, 2010; Baart, Stekelenburg, & Vroomen, 2014; Besle et al., 2004; van Wassenhove et al., 2005). Many of these interactions occur so early they are unlikely to have been produced by feedback from putative integration areas, and instead suggest that multisensory information is fed directly into primary sensory processing areas (See Besle et al., 2008 for a review; See also Shahin et al., 2018). In fact there is evidence of visual information modulating auditory processing in the brainstem (e.g. Musacchia, Sams, Nicol, & Kraus, 2006). It would seem unlikely that Samuel and Lieblich's (2014) "linguistic-perceptual" dissociation begins earlier than this.

Relatedly, there is evidence that visual speech can even influence the functioning of the cochlea (otoacoustic emissions; Namasivayam, Wong, Sharma & van Lieshout, 2015). While this process is almost certainly the result of feedback from higher processing areas, it still poses a formidable challenge for the Samuel and Lieblich (2014) account. By influencing the functions of the ear, multisensory context is influencing the

auditory information that is initially acquired by the perceiver. In this sense, the linguistic information contained in auditory speech is influenced by multisensory context at the sensory organ. Thus in order for it to be viable, the Samuel and Lieblich (2014) account would need to explain how linguistic processing can proceed without multisensory influences, including the multisensory influences that affect how the *ear* acquires speech.

It is also arguable, that the empirical evidence in support of the Samuel and Lieblich (2014) account is not particularly strong. To briefly re-state these selective adaptation findings; McGurk adaptors produce selective adaptation to the unperceived auditory stimulus (e.g. see Roberts & Summerfield, 1981; Saldana & Rosenblum, 1994), while phonemic restoration (Samuel, 1997) and Ganong stimuli (Samuel, 2001; Samuel & Frost, 2015) support selective adaptation to a segment that is perceived but not present in the stimulus. Thus it seems that selective adaptation follows perception when that perception is determined by lexical information, but not when it is determined by multisensory information.

However it is also possible that these findings reflect the fact that the multisensory selective adaptation studies have relied on the McGurk effect, in which *clear* auditory speech is *presented simultaneously* with *clear* and incongruent visual speech (McGurk & MacDonald, 1976). In contrast, the lexical context selective adaptation effects have been found with: a) the phonemic restoration effect (Samuel, 1997) in which the adapting phoneme is absent and replaced with noise; and b) the Ganong effect (Samuel, 2001; Samuel & Frost 2014) in which the adapting phoneme is acoustically ambiguous. In both of these cases, these lexical context effects have been

observed with stimuli that contain *unclear* (ambiguous) segments devoid of any simultaneous competing information.

Thus, the failure of multisensory context to influence selective adaptation coincides with the *presence of clear and conflicting* phonetic information, while the success of lexical context to influence selective adaptation is based on *unclear* phonetic information being embedded in a supportive context —*without conflicting* information present. It could be that these superficial stimulus distinctions account for the diverging effects, rather than any difference in the roles of multisensory and lexical context information.

In fact, Samuel and Lieblich (2014) acknowledge this possibility, but argue that more than phonetic ambiguity must account for the lexically driven selective adaptation effects. They point to a series of studies that approximate the selective adaptation procedures and that also used an audio-visual stimulus in which the audio channel was phonetically ambiguous (as in the Ganong effect) but perceptually guided by an accompanying clear visual stimulus (Baart & Vroomen, 2010; Vroomen & Baart, 2009; Keetels, Pecoraro, & Vroomen, 2015; Bertelson, Vroomen, & De Gelder, 2003; Vroomen, van Linden, De Gelder, & Bertelson, 2007). Despite the methodological similarities to selective adaptation, and a stimulus composed of an acoustically ambiguous segment, these researchers failed to find adaptation to the perceived audio-visual stimulus (Vroomen et al., 2007). Thus, Samuel and Lieblich (2014) argue that even in the context of phonetic ambiguity, multisensory information is unable to drive selective adaptation.

However, it can also be argued that the ambiguous auditory + clear visual speech stimuli used by Vroomen and colleagues (2007) are not truly comparable to the ambiguous auditory segments tested in a word context (i.e. Ganong type) as used by Samuel (2001; Samuel & Frost, 2014). This is because the Vroomen and colleagues' (2007) ambiguous auditory + clear visual speech stimulus still retains *conflicting* audio-visual information. In contrast, traditional (lexical) Ganong type stimuli do not contain *conflicting information*. For this reason, the present investigation will test if the dissociation between lexical and multisensory context effects on selective adaptation is eliminated when the multisensory context lacks conflicting information.

The Current Study

The present investigation was designed to compare the effects of lexical and visual context on selective adaptation using *comparable critical stimuli* to test both contexts. To achieve this, we exploit the phonemic restoration effect, in which auditory information is removed from the stimulus and replaced by noise, and thus lacks the condition of conflicting information (see above). The phonemic restoration method will be applied to both lexical and multisensory contexts.

In the following experiments we will compare selective adaptation effects induced by two different kinds of phonemic restoration stimuli: non-lexical multisensory phonemic restoration, and audio-only lexical phonemic restoration. The stimuli for both of these conditions originated as audio-visual recordings of a talker saying words with either a central /d/ or /b/ segment (e.g. "arma**d**illo" & "inhi**b**ition"; see also Samuel, 1997). These central /b/ and /d/ segments were removed from the auditory channel and

replaced with noise to produce phonemic restoration stimuli (e.g. Warren, 1970). The audio-only lexical phonemic restoration stimuli were made by removing the visual channel from these stimuli, while the non-lexical multisensory restoration stimuli retained the visual channel but removed the initial and final segments of the words to produce audio-visual speech-noise-speech bi-syllables. The critical question addressed by the following experiments is whether these lexical and multisensory restoration stimuli support comparable selective adaptation effects. *If, as Samuel and Lieblich (2014) propose, selective adaptation is sensitive to a linguistic process that is insulated from multisensory information, then selective adaptation will only occur for lexical, but not multisensory, phonemic restoration contexts*. *If, on the other hand, the process that drives selective adaptation is sensitive to multisensory information, then both multisensory and lexical phonemic restoration effects should produce selective adaptation effects.*

These predictions were tested in three experiments. Experiment 1 served as a control, establishing that our full words, with no replacing noise, support selective adaptation effects (See also Samuel, 1997). In Experiment 2, the adapting segments of the words from Experiment 1 were removed and replaced with signal-correlated-noise to produce phonemic restoration stimuli. Experiment 2 had three conditions; lexical phonemic restoration (audio-only words + noise), multisensory phonemic restoration (audio-visual bi-syllables + noise), and a non-restoration control condition (*audio-only* bi-syllables + noise). All the stimuli used in Experiment 2 were adapted from the stimuli in Experiment 1. The audio-only bi-syllables were extracted from the same stimuli used for the lexical and multisensory context conditions, and thus made an ideal control. In

Experiment 3 we replicated the procedures of Experiment 2 using a different replacing noise.

Experiment 1

Experiment 1 began by testing the selective adaptation produced by the full word stimuli (those without any replacing noise). This experiment provides us with an assessment of the adaptation effects that are driven by the acoustic speech information whereas the subsequent experiments will assess adaptation to *illusory* speech percepts.

Method

Participants

Forty (16 male) University of California, Riverside students participated in Experiment 1 for course credit. All subjects were native English speakers and reported normal hearing and normal or corrected to normal vision.

Materials

All stimuli in this experiment were derived from audio-video recordings of natural words and syllables produced by a 22-year-old female speaker. This speaker was a monolingual English speaker native to Southern California. All productions were articulated at a natural pace.

Test Continuum. During audio-video recording the model alternated between /da/ and /ba/ syllables, producing multiple exemplars of each. The best recordings of each syllable were used to generate the test continuum. The continuum was designed by synthetically interpolating the formant frequencies of the first three formants between the recorded produced /ba/ and /da/ syllables using a script available from

(http://www.mattwinn.com/praat.html). The natural syllables served as endpoints to the continuum and had the values of (Ba: F1: 1425hz; F2: 2491hz; F3: 5436) and (Da: F1: 790hz; F2: 26501hz; F3: 3876).

Adaptation Stimuli. The adaptation stimuli consisted of the audio channel of audio-visual recordings of words of three or more syllables with /d/ or /b/ segments in the middle of the utterance. These words were "Recon**d**ition," "Arma**d**illo," "Confi**d**ential," "Aca**d**emic," "Psyche**d**elic," "Canni**b**al," "Alpha**b**et," "Cere**b**ellum," "Cari**bb**ean," and "Inhi**b**ition" (These were the same /b/ words used by Samuel 1997; the only exception being that we substituted "Cannibal" for "Exhibition" as we were concerned that "Exhibition" may be too similar to "Inhibition").

Procedure

Each subject was assigned to either the /b/ (20 subjects) adaptor or the /d/ (20 subjects) adaptor condition (Dias, Cook, & Rosenblum, 2016). In the first part of the experiment, subjects made their initial baseline judgments of the tokens in the ba-da test continuum. During this portion of the experiment, subjects listened to the test items, one at a time, and for each item, reported either /da/ or /ba/ by pressing one of two labeled buttons on a computer keyboard. The test items were presented to the subjects in a random order for 44 complete cycles of 8 continuum items (Eimas & Corbit, 1973; Samuel, 1986; Vroomen et al., 2007).

Following the baseline measurement, the experiment alternated between two subject tasks (Samuel, 1997). In the first task, subjects listened to a continuous stream of the adaptor word stimuli presented in a random order at a rate of approximately one word

per 1.5 seconds (the word length influenced the trial to trial duration). The primary instruction to subjects in this phase of the experiment was to listen to the auditory stimuli. Additionally, during this phase of the experiment, a white dot was displayed on the screen during a randomly selected 25% of the adapting words. Subjects were instructed to press the spacebar on the computer keyboard when they saw this dot. The purpose of this dot monitoring task was for consistency with Experiment 2 in which a similar methodology was used to encourage subjects to attend to the visual component of the adaptors.

The content of the adaptation stream depended on the condition—/d/ or /b/ segment adaptation—to which the subject was assigned. Subjects in the /d/ segment condition heard adapting words containing /d/ segments (e.g. "Recon**d**ition"), while subjects in the /b/ segment condition heard adapting words containing /b/ segments (e.g. "Inhi**b**ition"). In both conditions, the ordering of words in the adaptation stream was random without replacement.

Following adaptation, subjects were asked to identify all eight test continuum syllables presented in a random order. Subjects indicated their responses by pressing buttons labeled "Ba" or "Da." This portion of the experiment was identical to the baseline measure except that it consisted of only a single cycle of the test-continuum (Samuel, 1997).

The first adaptation sequence included 60 adaptor words, whereas all following adaptation sequences consisted of 40 adaptor words (Samuel, 1997). A total of 44 adaptation blocks and subsequent test-continuum identifications were included in this

experiment, approximating about 70 minutes in total duration (Eimas & Corbit, 1973; Samuel, 1986; Vroomen et al., 2007). A summary of this procedure is provided in Figure 2.1a.

A research assistant provided all instructions verbally, and these instructions were also printed as text and presented on the computer screen during the experiment. Instructions were administered at the start of the experiment and again before the first adaptation phase began.

Results

We began by tabulating the proportion of /ba/ identifications during the baseline and adaptation blocks. As can be seen in Figure 2.2, these /b/ and /d/ full word adaptors produced opposing identification shifts between the baseline and post adaptation (test) blocks. Similar to what Samuel (1997) reports, it appears that the identification shift was larger in the /d/ adaptation condition than in the /b/ adaptation condition.

Next we tested if these shifts constituted a significant selective adaptation. Samuel (1997; see also Samuel, 2001; Samuel & Frost, 2014; Samuel & Lieblich, 2014) measured selective adaptation by comparing the shift in continuum identifications for the middle four continuum items across the /b/ and /d/ adaptor conditions. We found a significant difference ($t[38] = 3.161$, $p = .002$, $r = .456$ [2 tailed][5]). These results replicate

[5] As selective adaptation is characterized by a change in the phoneme boundary it is common to restrict statistical tests to the middle of the test continuum where the boundary is located (See Samuel & Lieblich, 2014). In general measuring adaptation using the middle items is more sensitive then testing the full continuum, but the two comparisons should be similar. For the interested reader, this was the case with our data, the comparison of the full continuum was $t[38] = 2.544$, $p = .008$, $r = .381$ [2 tailed]) similar to what we found for the test of the middle items.

the results reported by Samuel (1997) and validate that our full word stimuli can support selective adaptation.

Experiment 2

Experiment 2 investigated if the adaptor stimuli of Experiment 1 would continue to support selective adaptation when the critical /b/ and /d/ segments were replaced by noise, in both audio-visual and lexical contexts. This experiment included three conditions; audio-visual bi-syllables, audio-only words, and audio-only bi-syllables. In each of these stimuli types, noise replaced the adapting /b/ or /d/ segments.

The audio-only words provided lexical, but not multisensory, context that was expected to support the classic phonemically-restored adaptation effects (Samuel, 1997). The audio-visual bi-syllables provided multisensory, but not lexical, context and were also expected to produce phonemic restoration (Abbott & Shahin, 2018). The question facing this experiment is whether these multisensory restoration effects would, like lexical restoration effects, produce selective adaptation. The audio-only bi-syllables provided neither lexical or multisensory context and were thus not expected to support phonemically restored selective adaptation. The stimuli for all three conditions were designed from the same audio-visual recordings making them directly comparable.

If selective adaptation is sensitive to a linguistic process that is insensitive to multisensory information, then selective adaptation will only occur for lexical, but not multisensory, phonemic restoration contexts. *If, on the other hand, the process that drives selective adaptation is sensitive to multisensory information, then both multisensory and lexical phonemic restoration effects should produce selective adaptation effects.*

Method

Participants

One hundred nineteen (77 male) University of California, Riverside students participated in Experiment 2 for course credit. Thirty-nine participants were assigned to the words with replacing noise condition (19 in the /b/ replaced), forty to the audio-visual bi-syllable condition (20 in the /b/ replaced condition), and forty in the audio-only bi-syllable condition (20 in the /b/ replaced condition) (Dias et al., 2016). All subjects were native English speakers and reported normal hearing and normal or corrected to normal vision.

Materials

The materials for this experiment consisted of the ba-da continuum used in Experiment 1 and the audio-only /b/ and /d/ words with replacing noise, audio-visual and audio-only bi-syllables with replacing noise that are described below (see also Figure 2.1b).

Adaptation Stimuli. The adaptation stimuli were created in two phases: replacing the critical adapting /b/ and /d/ segments with noise and then removing the unwanted contextual information to form our three stimulus conditions. Recall that Experiment 1 presented audio-only words that were extracted from audio-visual recordings. Using those original audio-visual recordings, we removed the /b/ and /d/ segments from the auditory channel. Next, for each word, we generated a white noise segment that retained

the intensity profile of the deleted /b/ or /d/ segment (i.e. signal-correlated-noise; Samuel, 1997; see also Figure 2.1b). These signal-correlated-noise segments were then inserted into the audio files for each corresponding word at the point where the removed /b/ or /d/ segment had originally been (See Figure 2.1b). Thus these /b/ and /d/ correlated noise segments replaced the real /b/ and /d/ segments[6].

Following the insertion of the noise segments, we edited these audio-visual words to create lexical and multisensory phonemic restoration context stimuli (and non-restoration control stimuli). The lexical phonemic restoration stimuli were created by removing the visual channel from the words resulting in audio-only words with noise replacing the /b/ or /d/ segments. These stimuli retained the lexical information specifying the identity of the segment replaced by noise and are comparable to those used by Samuel (1997). Accordingly, these stimuli should support the classic phonemically-restored adaptation effects.

The multisensory restoration stimuli were created by removing the initial and final segments of each word, so that only the replacing noise and the adjacent vowels remained (i.e. for each word the bi-syllable is indicated by the bolded segments shown here: Rec**o#i**tion," "Arm**a#i**llo," "Conf**i#e**ntial," "Ac**a#e**mic," "Psych**e#e**lic," "Cann**i#a**l," "Alph**a#e**t," "Ce**r#e**llum," "Car**i#e**an," and "Inh**i#i**tion"; see also Figure 2.1b). This editing produced audio-visual bi-syllables with noise. The video of the bi-syllable articulation was retained, and two images corresponding to the start and the end of the auditory bi-syllable respectively were added. The silent still images were presented for

[6] In addition to the adapting consonant, sections of the adjacent vowels were also removed and replaced with noise. This was done to remove co-articulation from the consonant.

durations that made the bi-syllable stimuli correspond to the duration of the original full word utterances from which they were derived. The resulting stimulus for each adaptor consisted of: 1) a silent still image of the speaker's articulatory position leading into 2) the synchronized audio and dynamic visual components of the critical bi-syllable (with signal-correlated-noise replacing the critical /b/ or /d/ segment), and then: 3) a silent still image of the speaker's ending articulation of the bi-syllable. Thus these audio-visual stimuli lacked the lexical context present in the audio-only words with noise, but instead had visual information specifying the identity of the noise-replaced segment. By lacking any conflicting crossmodal information as in the McGurk effect, these stimuli provide a more analogous test of contextual information on phonemic restoration effect.

Finally, the non-restoration control stimuli used these same bi-syllables but removed the visual channel. Being audio-only bi-syllables, these stimuli lacked both lexical and multisensory information and were not expected to support phonemic restoration-based adaptation effects.

Procedure

With the exception of the stimuli, the procedure of this experiment was identical to what was described for Experiment 1.

Each subject was assigned to one of the adaptor conditions (Dias et al., 2016). In the first part of the experiment, subjects made their initial baseline judgments of the tokens in the ba-da test continuum. During this portion of the experiment, subjects listened to the test items, one at a time, and for each item, reported either /da/ or /ba/ by pressing one of two labeled buttons on a computer keyboard. The test items were

presented to the subjects in a random order for 44 complete cycles of 8 continuum items (Eimas & Corbit, 1973; Samuel, 1986).

Following the baseline measurement, the experiment alternated between subjects listening/watching a continuous stream of the adaptor stimuli for their condition (each presented in a random order at a rate of approximately one item per 1.5 seconds). During this phase of the experiment, a white dot was displayed on the screen during a randomly selected 25% of the adapting words. Subjects were instructed to press the spacebar on the computer keyboard when they saw this dot. The purpose of this dot monitoring task was to encourage subjects in the AV condition to attend to the visual component of the adaptors.

Results

As was done for Experiment 1, we began our analysis by tabulating the proportion of /ba/ identifications on the test continuum at baseline and following adaptation. The condition means are presented in Figures 2.3-6. The results of our inferential analysis are detailed below.

Lexical Phonemic Restoration Adaptation

The /b/ and /d/ replaced contexts produced clearly opposing adaptation effects. This pattern was the result of a small (<1%) decline in /ba/ identifications for the /b/-replaced stimuli and a more pronounced increase in /ba/ identifications for the /d/-replaced stimuli (See Figure 2.3). The identification shift difference between /b/ and /d/ contexts (t[38] = 2.227, p = .032, r = .344 [2 tailed]) was statistically significant, demonstrating that these conditions did in fact produce selective adaptation (see Samuel

1997). This result replicates the primary finding reported by Samuel (1997); lexically based phonemic restoration appears to support selective adaptation.

Multisensory Phonemic Restoration Adaptation. The audio-visual /b/-replaced bi-syllables produced a small (.5%) *increase* in /ba/ identifications, an effect that is notably smaller than what was found for the /d/ replaced condition (see Figure 2.4). This pattern is also notable for being in the opposite direction from what was found for the full word /b/ and lexical context /b/ noise-replaced conditions however, it was not a significant shift from baseline (t[19] =.252, p = .804, r = .058 [2 tailed]) nor was it different from the /b/ conditions of either the Experiment 1 or the words with noise conditions discussed above (with replacing noise: t[37] = 0.349, p = .365, r = .057 [2 tailed]; no replacing noise: t[38] =.638, p = .264, r = .103 [2 tailed]).

As was done with the full words (Experiment 1) and the words with noise conditions, we next compared the identification shifts across the /b/ and /d/ conditions to determine if these audio-visual bi-syllables with noise produced multisensory phonemic restoration selective adaptation effects. This test showed a significant difference (t[38] = 2.85, p = .006, r = .42 [2 tailed]) demonstrating that our audio-visual contexts were producing the expected phonetically opposing adaptation effects (Samuel, 1997). This is a critical finding, it supports that multisensory information can produce selective adaptation effects, a result that challenges the account offered by Samuel and Lieblich (2014). The implications of this finding will be discussed below.

Comparing Lexical and Multisensory Mediated Adaptation

The above tests demonstrate that both lexical and multisensory contexts can support selective adaptation. We next tried to compare these effects to determine if selective adaptation was more sensitive to either lexical or multisensory information.

We began by assessing the baseline-test differences for each condition. We found that the lexical restoration /b/ context condition was not significant (t[18] = -0.248, p = .807, r = .058 [2 tailed]) but there was an effect of the lexical /d/ context (t[19] = 3.828, p < .001, r = .66 [2 tailed]). Mirroring these results, we found that the multisensory restoration /b/ context did not produce a reliable identification shift (see above) but the /d/ -replaced condition did (t[19] = 4.615, p < .001, r = .727 [2 tailed]).

We followed this test with an analysis to determine if multisensory and lexical contexts produced different selective adaptation effects across segments. This test compared the identification shifts for the /d/ replaced conditions across the lexical and multisensory contexts, but found no effect (t[38] =.351, p = .364, r = .057 [1 tailed]; see above for comparable test of the /b/ replaced conditions). Thus it seems that the multisensory and lexical conditions are comparable; the phonemic restoration effect on selective adaptation is stronger in /d/ replaced contexts than /b/ replaced contexts (see also the results of Samuel, 1997), but is not differently affected by lexical vs. multisensory context.

Non-Restoration Adaptation

A critical question facing these results is whether these adaptation effects can be attributed to phonemic restoration, that is: are the lexical or multisensory contexts driving

adaptation of the noise replaced segments, or are these effects driven by other, non-restoration, factors? To answer this question we began by examining the selective adaptation effects produced by our non-restoration stimuli: the audio-only bi-syllables with replacing noise. As stated, these stimuli were not expected to produce phonemic restoration, and any adaptation effect observed with these stimuli will necessarily be driven by the acoustic information contained in them.

Surprisingly, we found that both the /b/ replaced and /d/ replaced audio-only bi-syllables produced shifts in the direction of /d/ adaptation (see Figure 2.5). Samuel (1997) found a similar uniform shift pattern for noise-replaced segments in nonword stimuli. Samuel (1997) attributed this effect to a "labeling drift" for the participants (p. 11). This seems like the best explanation for our current results.

However, it is worth noting that these shifts *were* found to be significantly different from one another ($t[38] = 2.216$, $p = .033$, $r = .338$ [2 tailed]), suggesting that somehow the noise replaced /b/ and /d/ information was still influencing selective adaptation. As with our other conditions, this effect is driven by the larger shift in the /d/ replaced condition (see Figure 2.5). This is a surprising result; lacking both lexical and visual contextual information, these stimuli should not have supported phonemic restoration. Yet somehow these stimuli seem to be producing similar selective adaptation effects to what is observed in our restoration conditions.

Comparing Restoration and Non-Restoration Adaptation

As stated, we found a surprising and significant difference in the identification functions of the /b/ and /d/ replaced audio-only bi-syllables. This difference across /b/ and

/d/ replaced conditions is the essential evidence for restoration effects on selective adaptation. That we find this same difference with the audio-only bi-syllables begs the question of whether this test in our other conditions truly demonstrates *restoration* adaptation and not adaptation to the retained acoustic stimulus (the noise-replaced audio-only bi-syllable was present in all three contexts). Accordingly, we ran a series of tests comparing the identification shifts from the lexical (words with replacing noise) and multisensory (audio-visual bi-syllables with replacing noise) restoration conditions to the shifts from the non-restoration condition (audio-only bi-syllable with replacing noise).

Lexical Phonemic Restoration vs. Non-Restoration Comparisons. The baseline to test identification shifts from /b/ and /d/ replaced *lexical* context conditions were compared to the identification shifts from the corresponding audio-only bi-syllable conditions. Neither comparison was found to be significant (/b/ contexts: $t[37] = 1.823$, $p = .122$, $r = .191$ [1 tailed]; /d/ contexts: $t[38] = 0.798$, $p = .215$, $r = .128$ [1 tailed]; see Figure 2.6), indicating that there was no effect of lexically supported phonemic restoration that was not accounted for by the adaptation of the bi-syllables with noise. In other words, the provocative pattern of lexically supported selective adaptation reported above (see also Samuel, 1997) may have been driven by the acoustic information retained in the stimulus, not the top down lexical information.

Multisensory Phonemic Restoration vs. Non-Restoration Comparisons. The baseline to test identification shifts from /b/ and /d/ replaced *multisensory* context conditions were compared to the identification shifts from the corresponding audio-only bi-syllable conditions. However, similar to what was found for the lexical context

conditions, these audio-visual bi-syllables also did not produce adaptation effects significantly different from the effects produced by the audio-only bi-syllables (/b/-contexts: $t[38] = 1.092$, $p = .141$, $r = .175$ [1 tailed]; /d/-contexts: $t[38] = 0.528$, $p = .3$, $r = .085$ [1 tailed]; see Figure 2.6b). Thus, like what was found for the lexical context conditions, the audio-visual bi-syllables produced phonetically contrastive effects, but those effects may have been driven by the underlying acoustic information.

Discussion

The main goal of Experiment 2 was to replicate the original finding of lexically mediated selective adaptation (Samuel, 1997) and to compare this effect to the effects of multisensory mediated selective adaptation. Samuel (1997) operationally defined selective adaptation as a significant difference between the identification shifts observed across the /b/ and /d/ conditions. In finding that lexical context did induce this type of adaptation effect, we were able to replicate the results of Samuel (1997).

Even the magnitude of this difference is comparable between studies: Samuel (1997) reports a difference between conditions of 8.1%, just as we find a 8.1% difference (see Figure 2.3). In a further similarity to Samuel (1997), our phonetic contrast was driven by the larger effect of /d/ replaced stimuli, accounting for a 6.1% shift in Samuel (1997) and a 7.3% shift in our own study. Based on these data, it would seem that we did replicate the classic lexical selective adaptation effect.

Moreover, based on the comparison of /b/ and /d/ replaced audio-visual bi-syllable conditions, it seems that we have extended this original finding to multisensory contexts. This apparent equal access to the processes that drive selective adaptation for

lexical and multisensory contexts is antithetical to the account put forth by Samuel and Lieblich (2014). This point will be elaborated later.

However, another interesting result from Experiment 2 is that this effect seems to be driven by acoustic information retained in all stimuli. This is indicated by the fact that the audio-only bi-syllables with the replacing noise produce the same opposing /b/ vs. /d/ identification shifts as those observed in the lexical and audio-visual context conditions. That these bi-syllables were present in all conditions challenges any assertion that the effects in these conditions can be attributed to contextual information.

More than simply being a challenge to the present investigation, these results may have implications for the broader selective adaptation literature. Specifically, as noted above, our lexical context conditions closely matched Samuel (1997), both in design and in the resulting effects. Thus it may be possible that the lexically mediated effects reported by Samuel (1997) are likewise partly dependent on the remaining acoustic information in the signal and not just the lexical context, as such. This is especially concerning, as Samuel (1997) did *not* include a bi-syllable control condition in his study.

Instead Samuel (1997) *did* include a *non-lexical* control condition (noise in nonword contexts) that failed to produce a contrasting /b/ vs. /d/ replaced identification, as he predicted. However, this condition embedded the noise in *nonwords*. Thus it is possible that this nonword context *interfered* with the effects of the replacing noise that we observe here.

In this sense, the bi-syllables with replacing noise used in this current investigation are a better control in that they are the *exact* same stimuli that were

presented in the restoration conditions. Thus it is possible that the lexically mediated effects on selective adaptation reported by Samuel (1997) are dependent on the remaining acoustic information in the segment.

Experiment 3

That both the lexical and multisensory restoration conditions produced adaptation effects that were indistinguishable from the audio-only bi-syllables is concerning. While it is unclear what remaining acoustic structure in the control stimuli may have induced adaptation, a possible candidate is the signal-correlated nature of the replacing noise. It is known that selective adaptation can be driven by even rudimentary phonetic features, as stimuli such as white noise can produce selective adaptation on a fricative continuum (e.g. Samuel & Newport, 1979). While we were not using a fricative continuum, signal-correlated-noise is known to bolster phonemic restoration effects relative to other replacing sound, putatively because of its similarity to the replaced speech segment (Samuel, 1981). This is likely related to the fact that signal-correlated-noise can also carry some rudimentary acoustic phonetic information as shown by above chance performance in phoneme identification tasks (Shannon et al., 1995).

It is therefore possible that the selective adaptation effects noted in the previous experiment (and, possibly, Samuel's [1997] original study) were driven by this acoustic phonetic information in the replacing noise rather than phonemic restoration, as such. To investigate if Experiment 2's selective adaptation effects were driven by the acoustic phonetic information in the replacing noise, Experiment 3 replicated Experiment 2, but instead used *fixed amplitude* white noise as the replacing noise. Fixed amplitude noise

uses the same carrier signal as signal-correlated-noise. The key difference between these two stimuli is that unlike signal-correlated-noise, the temporal intensity profile of fixed amplitude noise does not correspond to the replaced speech signal, or anything (see Figure 2.1b). In contrast, the intensity profile of signal-correlated noise is *correlated* with the intensity profile of the speech signal it replaces. Importantly, this feature of signal-correlated-noise, which is absent from fixed amplitude noise, means that there are audible differences between the signal-correlated noise that replaced our /b/ segments and the signal-correlated-noise that replaced our /d/ segments (See Figure 2.1b). Fixed amplitude noise lacks this systematic correspondence to our /b/ and /d/ segments.

While fixed amplitude noise lacks much of the structure of signal correlated noise, it has been shown to support phonemic restoration (See Samuel, 1981). The question facing the current experiment is, will the phonemic restoration effects produced by fixed amplitude noise be sufficient to produce selective adaptation?

Method

Participants

One hundred fourteen (45 male) University of California, Riverside students participated in Experiment 3 for course credit. Thirty-seven participants were assigned to the words with replacing noise condition (20 in the /b/ replaced), thirty-seven to the audio-visual bi-syllable condition (17 in the /b/ replaced condition), and forty in the audio-only bi-syllable condition (20 in the /b/ replaced condition). All subjects were native English speakers and reported normal hearing and normal or corrected to normal vision.

Materials

The materials for this experiment consisted of the ba-da continuum and the audio-only /b/ and /d/ words with replacing noise, and audio-visual and audio-only bi-syllables with replacing noise that are described above. The replacing noise used in this experiment was fixed amplitude white noise of the same duration as the segment it replaced and scaled to the average intensity of the words (without noise) in which it was inserted (Samuel, 1981).

Procedure

With the exception of the stimuli, the procedure of this experiment was identical to what was described for Experiment 2. Briefly, participants first provided /ba/ vs. /da/ categorizations for 44 repetitions of the 8 continuum items, before going through 44 cycles of adaptation (exposure to adapting stimuli) and continuum categorizations. Participants were divided into lexical (audio-only words with noise) and multisensory (audio-visual bi-syllables with noise) restoration adaptation conditions, and non-restoration (audio-only bi-syllables with noise) conditions.

Results

Non-Restoration Adaptation

The central question of Experiment 3 was whether lexical and multisensory context effects on selective adaptation could occur without the support of signal-correlated-noise. Accordingly, we began our analysis with the audio-only noise-replaced bi-syllable conditions. Both the /b/ and /d/ replaced audio-only bi-syllables produced small shifts towards fewer /ba/ identifications at test (see Figure 2.7). Importantly, there

was no hint of opposing adaptation effects between the /b/ and /d/ replaced conditions (t[35] = 0.524, p = .604, r = .088 [2 tailed]). Recall that this comparison was the basis for the phonemic restoration selective adaptation effects reported in Experiment 2 and in Samuel (1997). Recall also that this comparison for the audio-only bi-syllables with signal-correlated-noise stimuli of Experiment 2 raised questions about the apparent "phonemic restoration" effects of that experiment. That this test for Experiment 3 is null means that significant effects found in the restoration conditions of this experiment will not be easily attributable these bi-syllables with noise.

Lexical Phonemic Restoration Adaptation

Having established that fixed amplitude replacing noise (in audio-only bi-syllable context) does not, on its own, produce the opposing identifications shifts between the /b/ and /d/ contexts, we next investigated whether adding lexical context would. Recall that Experiment 2 found that lexical context did produce selective adaptation, however, Experiment 2 also found that these effects were not different from the adaptation produced by a control condition, the audio-only bi-syllable with noise. Experiment 2 used signal-correlated-noise as the replacing noise for both the lexical context and bi-syllable conditions, while the current experiment uses *fixed amplitude* noise. The question addressed here is whether lexical context will continue to support phonemic restoration selective adaptation with this fixed amplitude replacing noise.

As with the audio-only bi-syllables, both the /b/ and /d/ replaced audio-only words produced small shifts towards fewer /ba/ identifications (see Figure 2.9). Furthermore, similar to what was found with the audio-only bi-syllables, there was no

difference in the identification shifts between the /b/ and /d/ replaced conditions (t[35] = 0.563, p = .577, r = .095 [2 tailed]). This is the first test of phonemic restoration selective adaptation using fixed amplitude noise. It seems that lexical context can only support phonemic restoration selective adaptation in the presence of the supportive bottom-up information provided by signal-correlated-noise.

We next ran an analysis to test if lexical context produced any change from the effect of the audio-only bi-syllables. However, we found no effect in either the /b/-replaced (t[37] = 0.805, p = .213, r = .135 [1 tailed]) or the /d/-replaced condition (t[37] = 0.912, p = .184, r = .152 [1 tailed]; see Figure 2.9). Based on these results, it is difficult to conclude that lexical context had any effect on selective adaptation. It seems that the only effects in the lexical context conditions of both Experiment 2 and the current experiment are driven by the acoustic information in the replacing noise; lexical context conditions do not differ from the control audio-only bi-syllables. It seems possible that lexical context effects on phonemic restoration selective adaptation are dependent on the bottom up supportive information from the acoustic signal, such as what is available in signal-correlated-noise (i.e. Experiment 2).

Audio-visual bi-syllables

The audio-visual /b/-replaced bi-syllables produced a non-significant increase in /ba/ identifications (t[19] = 1.311, p = .205, r = .288 [2 tailed]; see Figure 2.10). As with the results of Experiment 2, given that this shift was not significant, we continued on with our comparison of the shifts between the /b/ and /d/ replaced audio-visual bi-syllable conditions.

Unlike the results for the lexical context conditions, for the audio-visual bi-syllable condition there *was* in fact a difference between our /b/ and /d/ replaced conditions (t[35] = 2.718, p = .01, r = .417 [2 tailed]). This demonstrates that the multisensory context continued to support selective adaptation, even in the absence of the supportive acoustic information from signal-correlated-noise (See Figure 2.10c). That we failed to find a significant effect for this comparison using the audio-only bi-syllables from the present experiment suggests that this effect with the audio-*visual* bi-syllables *is* being driven by the multisensory contextual information. This is evidence for multisensory phonemic restoration selective adaptation. Furthermore, that the corresponding comparison was not significant for the lexical context conditions discussed above, suggests that these multisensory restoration effects are more reliable than are lexical effects. It seems that, unlike lexical context, multisensory context can continue to influence selective adaptation even in the absence of the supportive information afforded by signal-correlated-noise.

Comparing Lexical and Multisensory Mediated Adaptation

The primary motivation for this investigation was to determine whether multisensory and lexical contexts produce comparable selective adaptation effects. To test this question we next analyzed the difference between the identification shifts produced by our multisensory restoration (audio-visual bi-syllables with noise) to the shifts produced by the lexical restoration (audio-only words with noise) conditions. Here we found a reliable effect of the /d/-replaced conditions (t[32] = 2.594, p = .007, r = .417 [1 tailed]; Figure 2.11) indicating that the multisensory context condition produced a

larger adaptation effect than the lexical context condition for this segment. The /b/ replaced conditions were not found to be different; $t[38] = 0.299$, $p = .383$, $r = .048$ [1 tailed]. Thus, it seems that while the effects of multisensory context on selective adaptation are small, and more apparent in /d/ than /b/ contexts, they are larger and more reliable than the effects of lexical context. That is, contrary to what the account put forward by Samuel and Lieblich (2014) assumes, multisensory information appears to have *more* access to the processes that drive selective adaptation than does lexical information, based on the current methodology.

Discussion

There were several motives for running Experiment 3. One purpose of Experiment 3 was to replicate the principle finding of Experiment 2, that multisensory context could support phonemic restoration selective adaptation, when using fixed amplitude replacing noise. As the results of both Experiment 2 and 3 show shifts from the /d/ replaced audio-visual bi-syllables being different from the shifts from the /b/ replaced audio-visual bi-syllables, Experiment 3 was successful in this regard.

Another goal of Experiment 3 was to determine if the lexically mediated phonemic restoration effect on selective adaptation, first reported by Samuel (1997), would be found when replacing signal-correlated-noise with fixed-amplitude noise. Put simply, the results of the present experiment failed in this regard; no /b/ vs. /d/ identification differences were observed for the lexical context conditions of Experiment 3. Indeed, a post hoc test showed a significant decline in the identification shifts produced by the /d/ replaced by fixed amplitude noise relative to what was found for the signal-

correlated-noise replaced conditions (t[35] = 1.695, p = .049, r = .275 [1 tailed]; see also Figure 2.12a), that was absent from the corresponding test for the audio-visual bi-syllables (t[35] = -0.061, p = .476, r = .01 [1 tailed]; See Figure 12b[7]). It seems that signal-correlated-noise plays a role for lexical, but not multisensory, phonemic restoration selective adaptation.

An important finding of Experiment 3 is that, unlike Experiment 2, the audio-only bi-syllables did not produce the /b/ vs /d/ differences that are characteristic of selective adaptation. This means that the successful effects found for the audio-visual bi-syllables cannot be attributed to their audio channels. In contrast, the lexical context conditions seem to only produce the /b/ vs /d/ difference when the audio-only bi-syllables also do so, as they did in Experiment 2 (but not 3).

The chief difference between experiments 2 and 3 was the type of noise used to replace the /b/ and /d/ segments. Experiment 2 used signal-correlated-noise while Experiment 3 used fixed-amplitude noise. As their names imply these two types of noise differ in their relationship to speech; signal-correlated-noise replicates the speech acoustic speech intensity profile while fixed amplitude noise does not. That the signal-correlated-noise, without any lexical or multisensory context, produced phonemic restoration selective adaptation demonstrates how influential, even the minimal speech information retained in signal-correlated-noise can be.

[7] As across every experiment in this investigation (and indeed as was generally the case for Samuel, 1997) adaptation effects were characterized by larger shifts in the /d/ replaced conditions, thus these were the focus of our post-hoc tests. It is worth noting that the comparisons of the /b/ replaced conditions were null.

General Discussion

Across Experiments 2 and 3 we provide two comparisons of selective adaptation effects of lexical and multisensory contexts that were matched with respect to the acoustic support for /b/ and /d/. Experiment 2 compares selective adaptation from lexical and multisensory /b/ and /d/ restoration using signal correlated replacing noise, while Experiment 3 compares across conditions with fixed amplitude replacing noise.

Before discussing the effects of multisensory and lexical context, we should comment on the results of our /b/ replaced conditions. Across all of our experiments, the /d/ conditions always produced the larger shifts. This raises questions about the adaptation for the /b/ stimuli. In several instances, the /b/ context conditions produced effects that were actually in the direction of /d/ adaptation; however, in these instances the identification shift was always small and not significantly different from baseline. It should be noted that the results reported by Samuel (1997) also show a smaller adaptation effect for /b/ replaced conditions relative to /d/ replaced conditions. Thus it is possible that the effects reported here are more reflective of phonemically restored /b/ being a weak adaptor, than a limitation specific to our stimuli.

Next we should consider what our experiments indicate about the role of replacing noise in producing lexically and multisensory supported restoration selective adaptation. The results of Samuel (1997) have been taken to indicate that selective adaptation can be driven by a top-down effect of lexical processing of phonemic information. However, Experiment 2 and Experiment 3 used identical lexical contexts, the only difference between the stimuli was the replacing noise. The replacing noise of

Experiment 2, like the replacing noise of Samuel (1997), retained the intensity profile of the segments it replaced, while the replacing noise of Experiment 3 had no such correspondence to the replaced speech. That we only find lexically mediated selective adaptation effects for Experiment 2 but not Experiment 3 highlights this crucial role of the *bottom-up* information provided by the signal-correlated-noise to the putatively top-down lexical effects.

It is difficult to infer to what extent our findings apply to the results of Samuel (1997). Samuel (1997) did not include a fixed amplitude noise condition (but see Samuel [1981] who studied fixed amplitude noise in phonemic restoration for a different paradigm). Nor did he compare his lexical context adaptation effects to the adaptation effects from his non-lexical context (nonword) conditions. As these were two of the most important factors of the present study, direct comparisons between our findings and the results of Samuel (1997) are limited.

It should certainly be pointed out that in Samuel's (1997) non-lexical context condition, his identification shifts were in the opposite direction from ours. Based on this, rather than conclude that all lexical context effects on selective adaptation are dependent on signal-correlated noise, we conclude that lexical context effects are somewhat fragile and sensitive to idiosyncrasies of the stimuli which they are embedded.

Recall that Samuel (1997) operationally defined context driven selective adaptation effects as phonetically contrasting identification shifts. By this criteria we unequivocally find evidence that multisensory context can support adaptation effects. Unlike what we found for lexical contexts, these multisensory effects were present for

both the signal-correlated *and* fixed amplitude replacing noise in the audio-visual bi-syllables. In this sense, it seems that multisensory context is less dependent on the bottom-up information provided by signal-correlated-noise in producing selective adaptation effects.

An important motivation of Experiment 3, and of this entire investigation, was to determine if lexical and multisensory contexts produced comparable effects on selective adaptation. The above discussion indicates that these contexts are not comparable; the adaptation effects of lexical context seem less consistent than the effects of multisensory context, and more dependent of the presence of supporting acoustic phonetic information in the form of signal-correlated-noise. This conclusion is directly supported by our comparison of the /d/ replaced by fixed amplitude noise conditions which showed a significantly larger identification shift for the multisensory context over the lexical context condition.

Implications For Samuel and Lieblich (2014)

Over the last forty years a series of selective adaptation studies have shown that lexical, but not multisensory, illusions can drive selective adaptation. Based on the selective adaptation literature, and a pair of new experiments, Samuel and Lieblich (2014) argued that, relative to lexical information, multisensory information has a more limited and superficial effect on speech processing. These researchers argue that selective adaptation is one, of several, lines of converging evidence that supports their account in which speech supports separate linguistic and perceptual processes. The current study focuses on selective adaptation with respect to this account. We carefully compared

adaptation effects produced by lexical and multisensory information to address the critical theoretical question of whether either source of information is more fundamental to speech processing.

As noted above, the results of this investigation challenge the account offered by Samuel and Lieblich (2014). Based on these results we can conclude: 1) that multisensory context *can* produce selective adaptation effects; and 2) that this multisensory context can, in some ways, be more reliable than the lexical context effect on selective adaptation. Both of these conclusions challenge the account offered by Samuel and Lieblich (2014) which asserts that selective adaptation is driven by a process that is insensitive to multisensory information.

These conclusions converge with the results from two other investigations recently conducted in our lab. First, our lab has conducted a meta-analysis including results from an experiment conducted in our lab, as well as from the adaptation studies cited by Samuel and Lieblich (2014) (Dorsi et al., in prep). The focus of this analysis was to determine if the clear and conflicting information in McGurk adaptors causes a dilution of selective adaptation. Despite no single study finding a dilution effect for McGurk adaptors, we did find a significant dilution effect *across* studies (See also Dias, 2016). It seems that McGurk adaptors cause a small, but consistent, reduction in selective adaptation relative to audio-only adaptors.

Second, in a recent experiment we found that McGurk adaptors produce parallel and opposing auditory *and visual* selective adaptation effects (Dorsi, 2018 presentation). In other words, when selective adaptation was measured on an auditory continuum, the

auditory channel of the McGurk stimulus drove the effect, but when selective adaptation was measured on a visual continuum (see Dias, 2016) the visual (and perceived) channel of the McGurk stimulus drove the effect. Importantly we also found that the visual selective adaptation effect found for McGurk adaptors was significantly smaller than the visual adaptation effect of visual-only adaptors. This effectively replicates the results of our meta-analysis with visual selective adaptation, and suggests that the conflicting information in McGurk stimuli *does* impose a cost on selective adaptation. This result is another example of selective adaptation being sensitive to multisensory information.

Together with the experiments reported here, these two investigations suggest that selective adaptation is, in fact, sensitive to multisensory information. This challenges the account offered by Samuel and Lieblich (2014).

Evaluating Other Evidence For the Separate Perceptual and Linguistic Processes

Samuel and Lieblich (2014) argue that selective adaptation is not alone in demonstrating a dissociation between lexical and multisensory influences on speech processing. They discuss evidence for their proposed dissociation in: semantic priming, compensation for coarticulation, and neurophysiological data. We will discuss each of these points, and consider how supportive they are for the account offered by Samuel and Lieblich (2014).

Semantic Priming

The semantic priming data that Samuel and Lieblich (2014) cite in support of their account comes from Ostrand et al., (2016). The principle findings of this study were noted in the introduction section (and see Chapter 1), but to briefly summarize; Ostrand

et al., (2016) find that semantic priming was consistent with the auditory as opposed to the putatively perceived component of a McGurk word stimulus. For Samuel and Lieblich (2014) this finding reflects the linguistic process using the auditory signal to drive semantic priming independent of the audio-visual integrated percept. These researchers argue that the finding demonstrates a dissociation between how the stimulus was perceived and how it was linguistically processed.

However, it should be noted that the support for the Samuel and Lieblich (2014) account offered by Ostrand et al., (2016) is dependent on the assumption that semantic priming for McGurk words differed from the perception of those McGurk stimuli. We start our critique of the Samuel and Lieblich (2014) account by noting that this assumption is somewhat tenuous; McGurk stimuli do not always produce the McGurk effect, and Ostrand et al., (2016) did not measure how their McGurk stimuli were identified by their participants (see also Chapter 1 of this dissertation for a more extensive evaluation).

Furthermore, Chapter 1 demonstrates that while semantic priming can be consistent with the auditory-word of a McGurk stimulus, it is sometimes also consistent with the visual word. Critically, this research reports that whether semantic priming is consistent with the auditory or visual components of a McGurk word corresponds to how the McGurk word is *perceived*. Thus, contrary to what is assumed by the Samuel and Lieblich (2014) model, semantic priming shows that linguistic processing is consistent with, rather than independent from, perceptual processing.

Compensation for Coarticulation

Compensation for coarticulation is a perceptual phenomenon related to the continuous nature of speech, which results in temporal overlap of the articulation of adjacent segments; during speech a talker begins each word segment before completing the preceding segment. This "gestural overlap" (Fowler, 2010) affects the speech signal, for example, when isolated from the word "Bal**d**ing" the /d/ segment may sound more like a /g/ owing to its proximity to the preceding /l/. The perceptual system accommodates these artifacts of articulatory overlap, allowing listeners to "compensate for coarticulation" (e.g. Mann, 1980).

In another classic demonstration of compensation for coarticulation, more items from a /ta/-/ka/ continuum are identified as /ka/ when preceded by /s/, while more are identified as /ta/ when preceded by /ʃ/ (Mann & Repp, 1980). Relevant to the account of Samuel and Lieblich (2014) is work by Elman & McClelland (1988). This study produced a compensation for coarticulation effect that was driven by lexical context. That is, these authors induced compensation for coarticulation using a stimulus that was ambiguous between /s/ and /ʃ/ (/?/) that was appended to the end of either stimuli like "Fooli" (i.e. "Fooli**sh**"; or other segments of words that end with /ʃ/) or "Christma" (i.e. "Christma**s**"; or other segments of words ended that with /s/; see also the Ganong effect [Ganong, 1980]). In other words, a single /?/ segment can produce two contrasting compensation for coarticulation effects depending the lexical context in which it is

embedded (see also Magnuson, McMurrary, Tanenhaus, & Aslin, 2003; Samuel & Pitt, 2003).

In contrast, Vroomen and de Gelder (2001) failed to find a compensation effect using a multisensory context. These authors did find that when an auditory /?/ was dubbed with a visual /s/ it was identified as /s/ and when /?/ was dubbed with /ʃ/ is was identified as /ʃ/. However, despite these identification results, these authors failed to find a visually driven compensation for co-articulation effect. In this sense, the compensation for coarticulation effects seems to show a dissociation of lexical and multisensory effects on speech perception.

It is, however, worth noting that some research argues that non-lexical factors, such as transitional probabilities, can account for the apparent lexical effect on compensation for coarticulation (Pitt & McQueen, 1998; McQueen, 2003; McQueen, Jesse, & Norris, 2009). So whether or not compensation for coarticulation is lexically mediated, and therefore reflective of the linguistic processing postulated by Samuel and Lieblich (2014), remains uncertain.

It also worth noting that Vroomen and de Gelder (2001) is not the only investigation into multisensory influences on compensation for coarticulation. Fowler et al., (2000; see also Green & Norrix, 2001) have found visually-mediated compensation for coarticulation effects; a finding which does not fit well with the account offered by Samuel and Lieblich (2014). Similar to the debate surrounding the lexically mediated compensation for coarticulation findings, these visually mediated compensation for coarticulation effects have been challenged by authors suggesting that more general

learning principles can account for what appear to be multisensory context effects (Holt, Stephens, & Lotto, 2005; but see Fowler, 2006). As with the lexically mediated compensation of coarticulation findings, the debate concerning multisensory mediated compensation for coarticulation is extensive; however a recent meta-analysis (Viswanathan & Stephens, 2016) shows support for a multisensory role in compensation for coarticulation. Thus, in contrast to what is intimated by Samuel and Lieblich (2014), it seems that compensation for coarticulation can be sensitive to multisensory information.

Neurophysiological Findings

Samuel and Lieblich (2014) also cite neurophysiological data supporting their linguistic-perception dissociation. Samuel and Lieblich (2014) emphasize a series of findings concerning the audio-visual modulation of the auditory evoked N1 ERP (Besle et al., 2004; van Wassenhove et al., 2005, see also Stekelenburg & Vroomen, 2007). Samuel and Lieblich (2014) argue that, specifically the work by Vroomen and Stekelenburg (2007; 2010) shows that this audio-visual interaction is not limited to speech processing. For Samuel and Lieblich (2014) the audio-visual modulation of the N1 is an example of the brain engaging in a perceptual process that is separate from linguistic processing. This argument was solidified by Baart and Samuel (2015) who measured ERPs in response to audio-only, visual-only, or audio-visual words and nonwords. These authors report separate main effects for lexical context (words vs. nonwords) and multisensory contexts (audio-visual, audio-only, & visual-only speech), but no interaction between multisensory and lexical contexts. Consistent with the account

put forth by Samuel and Lieblich (2014), Baart and Samuel (2015) argue that their study demonstrates that the brain processes multisensory and lexical information in two separate neurological processes.

However, more recent research offers some counter evidence. Basirat, Brunelliere, and Hartsuiker (2018) used the word repetition effect in a recent EEG study to examine the effects of multisensory and linguistic processes. The word repetition effect is the finding that prior processing of words, but not nonwords, facilitates subsequent processing of those same words (e.g. subjects will identify a word faster the second time it is presented; Forbach, Stanners, & Hochhaus, 1974). The P200 ERP component is known to be modulated by word repetition (e.g. Almeida & Peoppel, 2013). However, Basirat et al., (2018) found that this repetition effect on the P200 interacted with multisensory context. Their results indicate that the multisensory information of audio-visual speech may facilitate word processing analogously to the facilitation provided by word repetition. This finding is a clear contrast to the interpretation offered by Samuel and Lieblich (2014), that the brain processes multisensory and linguistic information separately. Basirat et al., (2018) suggests that, at least in some circumstances, a single brain process may be responsible for both multisensory and linguistic information.

Conclusion

Samuel and Lieblich (2014) argue for separate linguistic and perceptual processes that have different functions for language processing. Under this account, the linguistic process is sensitive to lexical but not multisensory information, and this division seems to

occur at the very earliest stages of speech processing. Samuel and Leiblich (2014) support this account with findings from the selective adaptation literature that show selective adaptation can be driven by lexically supported, but not multisensory supported illusions. These authors further argue that these selective adaptation effects converge with findings from semantic priming, compensation for coarticulation, and neurophysiology.

In the preceding discussion we evaluated each of these findings and conclude that none offers conclusive evidence in support of the Samuel and Lieblich (2014) account. That none of this converging evidence is conclusive puts a sizable burden on selective adaptation for supporting their account. However, the experiments presented above show that, far from being insensitive to multisensory information, selective adaptation can be sensitive to, and in some cases is *more* sensitive to, multisensory information than it is to lexical information. This conclusion is difficult to reconcile with the account offered by Samuel and Lieblich (2014).

That selective adaptation seems more sensitive to multisensory, than lexical, context seems reasonable. Indeed, it is almost a certainty that multisensory context was informing our human ancestors before even the most basic languages existed. These are likely the reasons that multisensory information is integrated so early and why cross-sensory neuro-activity is so widespread (e.g. see Rosenblum et al., 2016 for a review). In light of this, it seems only reasonable that linguistic processing would be built around, not independent of, multisensory information. Indeed this appears to be born out in the literature discussed above concerning compensation for coarticulation, semantic priming

and of course selective adaptation: each case shows a primacy of multisensory information.

References

Abbott, N. T., & Shahin, A. J. (2018). Cross-modal phonetic encoding facilitates the McGurk illusion and phonemic restoration. *Journal of Neurophysiology*, *120*(6), 2988–3000. http://doi.org/10.1152/jn.00262.2018

Almeida, D., & Poeppel, D. (2013). Word-specific repetition effects revealed by MEG and the implications for lexical access. *Brain and language*, *127*(3), 497-509.

Baart, M., & Samuel, A. G. (2015). Turning a blind eye to the lexicon: ERPs show no cross-talk between lip-read and lexical context during speech sound processing. *Journal of Memory and Language*, *85*(July). http://doi.org/10.1016/j.jml.2015.06.00

Baart, M., Stekelenburg, J. J., & Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia*, *53*(1), 115–121. http://doi.org/10.1016/j.neuropsychologia.2013.11.011

Baart, M., & Vroomen, J. (2010). Phonetic recalibration does not depend on working memory. *Experimental Brain Research*, *203*(3), 575–582. http://doi.org/10.1007/s00221-010-2264-9

Basirat, A., Brunellière, A., & Hartsuiker, R. (2018). The role of audiovisual speech in the early stages of lexical processing as revealed by the ERP word repetition effect. *Language Learning*, *68*(June), 80–101. http://doi.org/10.1111/lang.12265

Bertelson, P., Vroomen, J., & De Gelder, B. (2003). Visual Recalibration of Auditory Speech Identification: A McGurk Aftereffect. *Psychological Science*, *14*(6), 592–597. http://doi.org/10.1046/j.0956-7976.2003.psci_1470.x

Besle, J., Fischer, C., Lecaignard, F., Bidet-Caulet, A., Lecaignard, F., Bertrand, O., & Giard, M. (2008). Visual activation and audiovisual interactions in the auditory cortex during speech perception : Intracranial recordings in humans. *The Journal of Neuroscience*, *28*(52), 14301–14310. http://doi.org/10.1523/JNEUROSCI.2875-08.2008

Besle, J., Fort, A., Delpuech, C., & Giard, M. H. (2004). Bimodal speech: Early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience*, *20*(8), 2225–2234. http://doi.org/10.1111/j.1460-9568.2004.03670.x

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., … David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science*, *276*(5312), 593–596. http://doi.org/10.1126/science.276.5312.593

Dias, J. W. (2016). *Crossmodal Influences in Selective Speech Adaptation* (Doctoral Dissertation). University of California, Riverside, Riverside, CA.

Dias, J. W., Cook, T. C., & Rosenblum, L. D. (2016). Influences of selective adaptation on perception of audiovisual speech. *Journal of Phonetics*, *56*, 75–84. http://doi.org/10.1016/j.wocn.2016.02.004

Dorsi, J. (in prepr). Dilution of selective adapatation effects by McGurk adaptors: A meta-analysis.

Dorsi, J. (2018, August). *Multisensory and lexical contexts in speech perception*. Talk given to Core for Advanced Magnetic Resonnce Imaging, at Baylor College of Medicine, Houston, TX.

Eimas, P. D., & Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, *4*, 99–109.

Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language*, *27*(2), 143–165. http://doi.org/10.1016/0749-596X(88)90071-X

Forbach, G. B., Stanners, R. F., & Hochhaus, L. (1974). Repetition and practice effects in a lexical decision task. *Memory & Cognition*, *2*(2), 337-339.

Fort, M., Kandel, S., Chipot, J., Savariaux, C., Granjon, L., & Spinelli, E. (2013). Seeing the initial articulatory gestures of a word triggers lexical access. *Language and Cognitive Processes*, *28*(8), 1–17. http://doi.org/10.1080/01690965.2012.701758

Fowler, C. A. (2006). Compensation for coarticulation reflects gesture perception, not spectral contrast. *Perception & Psychophysics*, *68*(2), 161–177. http://doi.org/10.3758/BF03193666

Fowler, C. A. (2010). Embodied, embedded language use. *Ecological Psychology*, *22*(4), 286–303. http://doi.org/10.1080/10407413.2010.517115

Fowler, C. A., Brown, J. M., & Mann, V. A. (2000). Contrast effects do not underlie effects of preceding liquids on stop-consonant identification by humans. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(3), 877–888. http://doi.org/10.1037//O096-1523.26.3.877

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology. Human Perception and Performance*, *6*(1), 110–125. http://doi.org/10.1037/0096-1523.6.1.110

Grant, K. W., & Seitz, P. F. P. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, *108*(3), 1197–1208. http://doi.org/10.1121/1.422512

Green, K. P., & Norrix, L. W. (2001). Perception of /r/ and /l/ in a stop cluster: Evidence of cross-modal context effects. *Journal of Experimental Psychology. Human Perception and Performance*, *27*(1), 166–177. http://doi.org/10.1037/0096-1523.27.1.166

Hirsh, I. J., Reynolds, E. G., & Joseph, M. (1954). Intelligibility of different speech materials. *The Journal of the Acoustical Society of America*, *26*(4), 530–538. http://doi.org/10.1121/1.1907370

Holt, L. L., Stephens, J. D. W., & Lotto, A. J. (2005). A critical evaluation of visually moderated phonetic context effects. *Perception & Psychophysics*, *67*(6), 1102–12. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16396017

Keetels, M., Pecoraro, M., & Vroomen, J. (2015). Recalibration of auditory phonemes by lipread speech is ear-specific. *Cognition*, 1–17.

Kim, J., Davis, C., & Krins, P. (2004). Amodal processing of visual speech as revealed by priming. *Cognition*, *93*(1). http://doi.org/10.1016/j.cognition.2003.11.003

Kleinschmidt, D. F., & Jaeger, T. F. (2016). Re-examining selective adaptation: Fatiguing feature detectors, or distributional learning? *Psychonomic Bulletin and Review*, *23*(3), 678–691. http://doi.org/10.3758/s13423-015-0943-z

Lüttke, C. S., Ekman, M., van Gerven, M. A. J., & de Lange, F. P. (2016). McGurk illusion recalibrates subsequent auditory perception. *Scientific Reports*, *6*(August), 32891. http://doi.org/10.1038/srep32891

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics*, *24*(3), 253–7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/704285

Magnuson, J. S., McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2003). Lexical effects on compensation for coarticulation: A tale of two systems? *Cognitive Science*, *27*(5), 801–805. http://doi.org/10.1016/S0364-0213(03)00067-3

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics*. http://doi.org/10.3758/BF03204884

Mann, V. a, & Repp, B. H. (1980). Influence of vocalic context on perception of the [zh]-[s] distinction. *Perception & Psychophysics*. http://doi.org/10.3758/BF03204377

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.

McQueen, J. M. (2003). The ghost of Christmas future: Didn't Scrooge learn to be good?

Commentary on Magnuson, McMurray, Tanenhaus, and Aslin (2003). *Cognitive Science*, *27*(5), 795–799. http://doi.org/10.1016/S0364-0213(03)00069-7

McQueen, J. M., Jesse, A., & Norris, D. (2009). No lexical–prelexical feedback during speech perception or: Is it time to stop playing those Christmas tapes? *Journal of Memory and Language*, *61*(1), 1–18. http://doi.org/10.1016/j.jml.2009.03.002

Miller, G. A, Heise, G. A, & Lighten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *The Journal of Experimental Psychology*, *41*(5), 329–335.

Musacchia, G., Sams, M., Nicol, T., & Kraus, N. (2006). Seeing speech affects acoustic information processing in the human brainstem. *Experimental Brain Research*, *168*(1–2), 1–10. http://doi.org/10.1007/s00221-005-0071-5

Namasivayam, A. K., Yiu, W., & Wong, S. (2015). Visual speech gestures modulate efferent auditory system, *14*(1), 73–83. http://doi.org/10.1142/S0219635215500016

Okada, K., Venezia, J. H., Matchin, W., Saberi, K., & Hickok, G. (2013). An fMRI study of audiovisual speech perception reveals multisensory interactions in auditory cortex. *PLoS ONE*, *8*(6), 1–8. http://doi.org/10.1371/journal.pone.0068959

Ostrand, R., Blumstein, S. E., Ferreira, V. S., & Morgan, J. L. (2016). What you see isn't always what you get: Auditory word signals trump consciously perceived words in lexical access. *Cognition*, *151*, 96–107. http://doi.org/10.1016/j.cognition.2016.02.019

Ostrand, R., Blumstein, S. E., & Morgan, J. L. (2011). When hearing lips and seeing voices becomes perceiving speech: Auditory-visual integration in lexical access. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *33*, 1376–1381. http://doi.org/10.1016/j.neuroimage.2010.12.063.Discrete

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I. P., Möttönen, R., Tarkiainen, A., & Sams, M. (2005). Primary auditory cortex activation by visual speech: An fMRI study at 3 T. *Neuroreport*, *16*(2), 125–128. http://doi.org/10.1097/00001756-200502080-00010

Pitt, M. A., & McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language*, *39*, 347–370. http://doi.org/10.1006/jmla.1998.2571

Roberts, M., & Summerfield, Q. (1981). Audiovisual presentation demonstrates that selective adaptation in speech perception is purely auditory. *Perception & Psychophysics*, *30*(4), 309–314. http://doi.org/10.3758/BF03206144

Rosenblum, L. D., Dias, J. W., & Dorsi, J. (2016). The supramodal brain: Implications for auditory perception. *Journal of Cognitive Psychology*, *5911*, 1–23. http://doi.org/10.1080/20445911.2016.1181691

Saldaña, H. M., & Rosenblum, L. D. (1994). Selective adaptation in speech perception using a compelling audiovisual adaptor. *The Journal of the Acoustical Society of America*, *95*(6), 3658–3661. http://doi.org/10.1121/1.409935

Samuel, A. G. (1981). The role of bottom-up confirmation in the phonemic restoration illusion. *Journal of Experimental Psychology. Human Perception and Performance*, *7*(5), 1124–31. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6457110

Samuel, A. G. (1986). Red herring detectors and speech perception: In defense of selective adaptation. *Cognitive Psychology*, *18*(4), 452–99. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3769426

Samuel, A. G. (1997). Lexical activation produces potent phonemic percepts. *Cognitive Psychology*, *127*(2), 97–127.

Samuel, A. G. (2001). Knowing a word affects the fundamental perception of the sounds within it. *Psychological Science*, *12*(4), 348–51.

Samuel, A. G., & Frost, R. (2015). Lexical support for phonetic perception during nonnative spoken word recognition. *Psychonomic Bulletin & Review*, (1970). http://doi.org/10.3758/s13423-015-0847-y

Samuel, A. G., & Lieblich, J. (2014). Visual speech acts differently than lexical context in supporting speech perception. *Journal of Experimental Psychology. Human Perception and Performance*, *40*(4), 1479–90. http://doi.org/10.1037/a0036656

Samuel, A. G., & Newport, E. L. (1979). Adaptation of speech by nonspeech: Evidence for complex acoustic cue detectors. *Journal of Experimental Psychology: Human Perception and Performance*, *5*(3), 563–578. http://doi.org/10.1037/h0078136

Samuel, A. G., & Pitt, M. A. (2003). Lexical activation (and other factors) can mediate compensation for coarticulation. *Journal of Memory and Language*, *48*, 416–434. http://doi.org/10.1016/S0749-596X(02)00514-4

Shahin, A. J., Backer, K. C., Rosenblum, L. D., & Kerlin, J. R. (2018). Neural mechanisms underlying cross-modal phonetic encoding. *The Journal of Neuroscience*, *38*(7), 1566–17. http://doi.org/10.1523/JNEUROSCI.1566-17.2017

Shannon, R. V, Zeng, F., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primaily temporal cues. *Science*, *270*(5234), 303–304.

Shigeno, S. (2002). Anchoring effects in audiovisual speech perception. *Journal of the Acoustical Society of America*, *111*(6), 2853–2861. http://doi.org/10.1121/1.1474446

Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience*, *19*(12), 1964–1973. http://doi.org/10.1162/jocn.2007.91213

Stekelenburg, J. J., & Vroomen, J. (2012). Electrophysiological correlates of predictive coding of auditory location in the perception of natural audiovisual events. *Frontiers in Integrative Neuroscience*, *6*, 26. http://doi.org/10.3389/fnint.2012.00026

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*(2), 212–215.

van Linden, S. (2007). Recalibration by auditory phoneme perception by lipread and lexical information (Doctoral thesis). Tilburg University, Tilburg, The Netherlands. ISBN:978-90–5335-122–2.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(4), 1181–1186. http://doi.org/10.1073/pnas.0408949102

Viswanathan, N., & Stephens, J. D. W. (2016). Compensation for visually specified coarticulation in liquid–stop contexts. *Attention, Perception, & Psychophysics*. http://doi.org/10.3758/s13414-016-1187-3

Vroomen, J., & Baart, M. (2009). Recalibration of phonetic categories by lipread speech: Measuring aftereffects after a 24-hour delay. *Language and Speech*, *52*(2–3), 341–350. http://doi.org/10.1177/0023830909103178

Vroomen, J., & de Gelder, B. (2001). Lipreading and the compensation for coarticulation mechanism Vroomen de Gelder 2001.pdf. *Language and Cognitive Processes*, *16*(5/6), 661–672.

Vroomen, J., van Linden, S., de Gelder, B., & Bertelson, P. (2007). Visual recalibration and selective adaptation in auditory-visual speech perception: Contrasting build-up courses. *Neuropsychologia*, *45*(3), 572–577. http://doi.org/10.1016/j.neuropsychologia.2006.01.031

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*(3917), 392–393. http://doi.org/10.1126/science.167.3917.392

Figure 2.1a

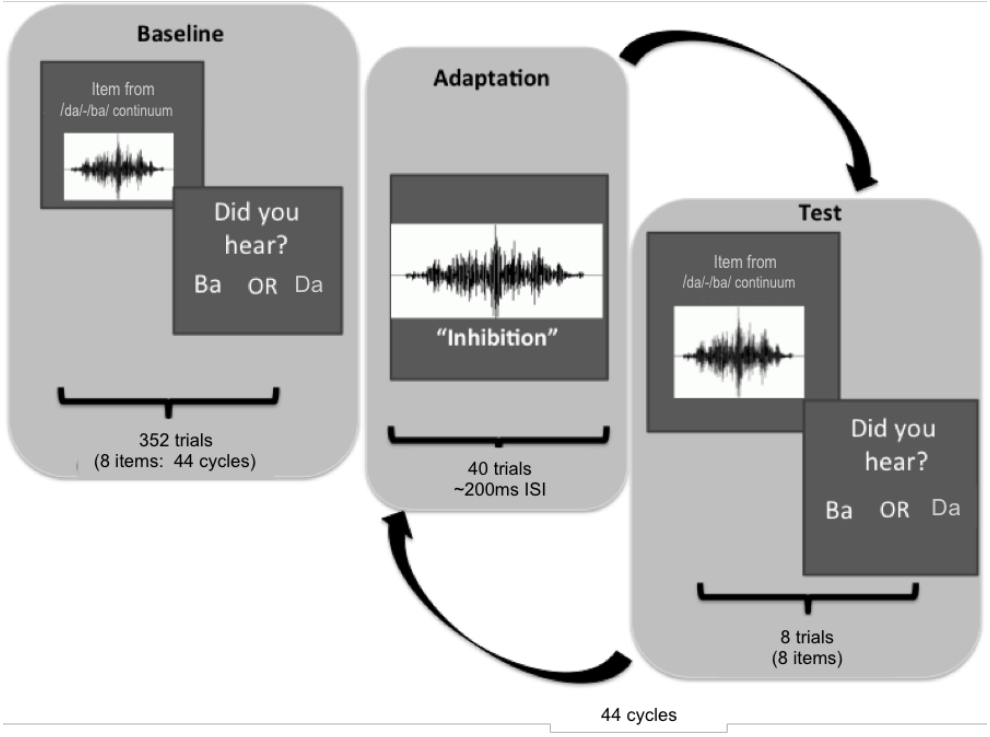


Figure 2.1a depicts the basic format the selective adaptation procedure for all three experiments.

Figure 2.1b

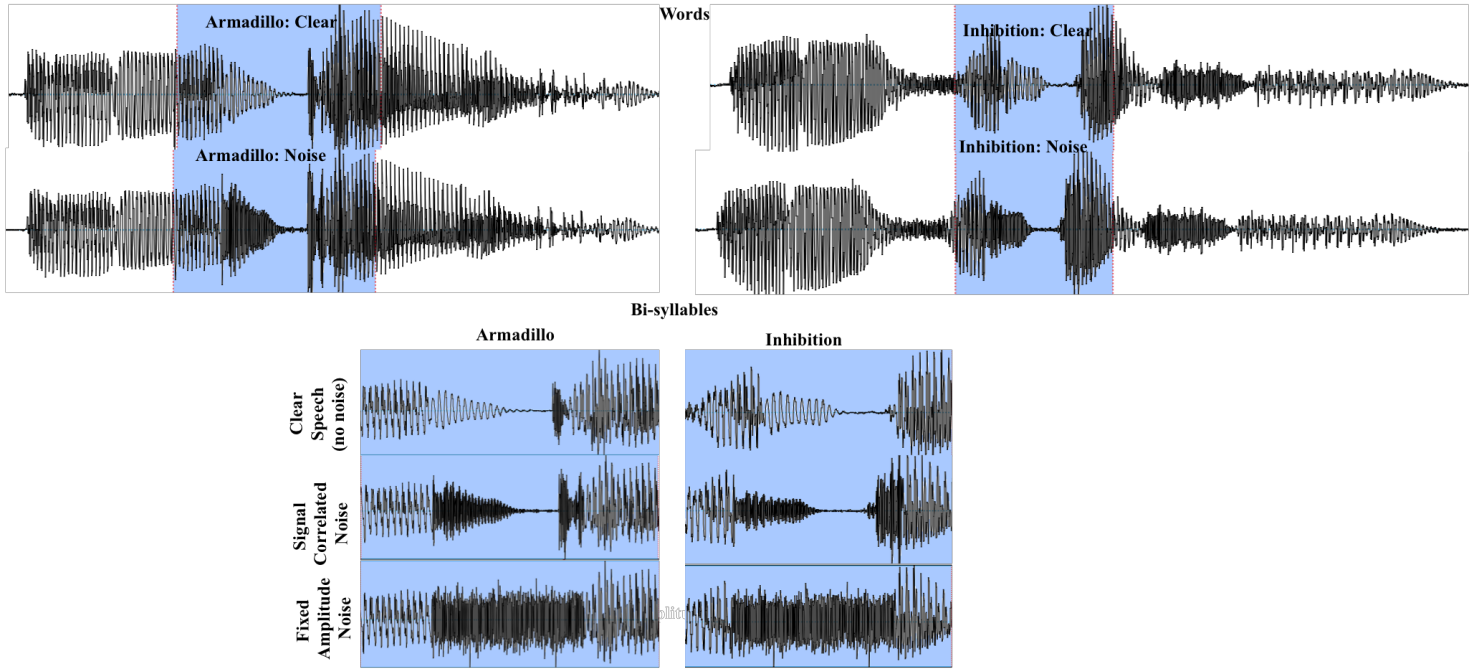
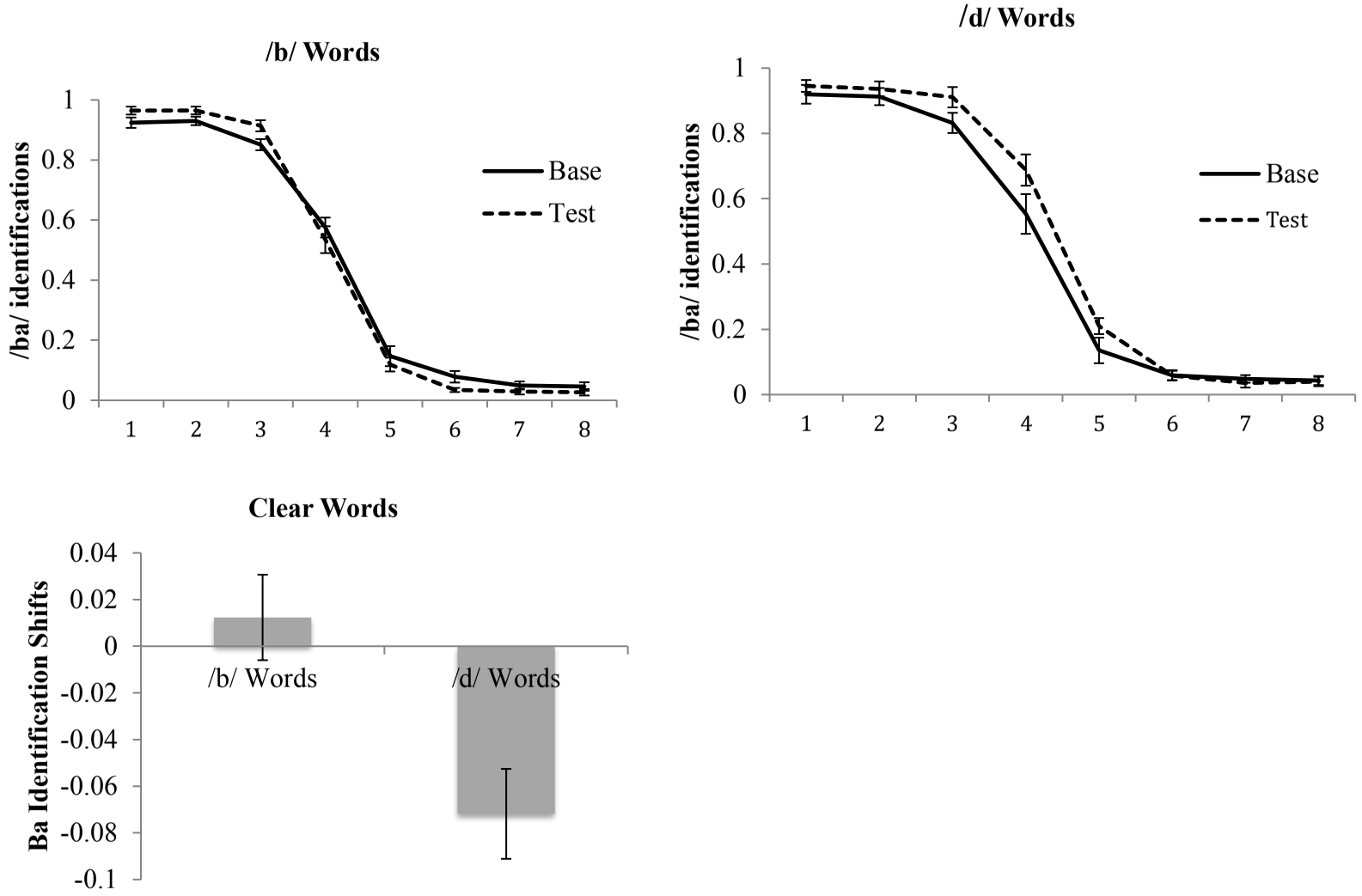


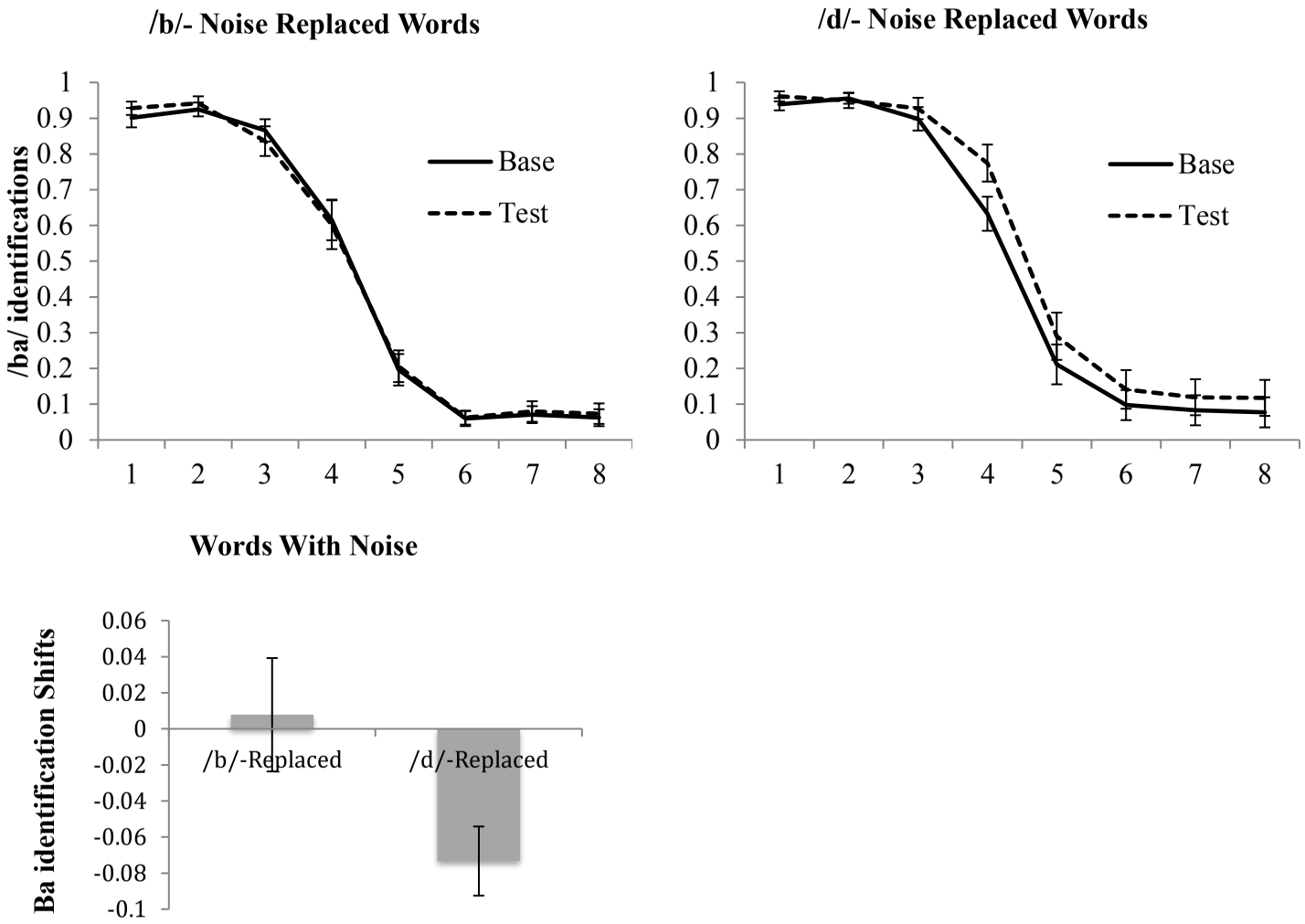
Figure 2.1b illustrates the auditory stimuli used in experiments 1 and 2. The top row shows the clear, no noise, speech "Armadillo" (left) and "Inhibition" (right) used in Experiment 1. The second row shows those same words, with the adapting /d/ and /b/ segments removed and replaced with signal-correlated-noise. Note that due to coarticulation, the replacing noise includes sections of the vowels adjacent to the adapting consonant. The sections highlighted in blue are the sections that were excised from the word context to be presented as bi-syllables. The third and fourth rows show enlargements of these sections. Note that the bi-syllables presented to participants always had replacing noise, the clear speech bi-syllables shown here are for comparison purposes only. The fifth row shows bi-syllables with non-signal correlated noise ("Fixed Amplitude Noise") which was used in place of signal-correlated-noise during Experiment 3.

Figure 2.2


/b/ Words
/ba/ identifications
1
0.8
0.6
0.4
0.2
0
1 2 3 4 5 6 7 8
Base
Test
/d/ Words
/ba/ identifications
1
0.8
0.6
0.4
0.2
0
1 2 3 4 5 6 7 8
Base
Test
Clear Words
Ba Identification Shifts
0.04
0.02
0
-0.02
-0.04
-0.06
-0.08
-0.1
/b/ Words
/d/ Words

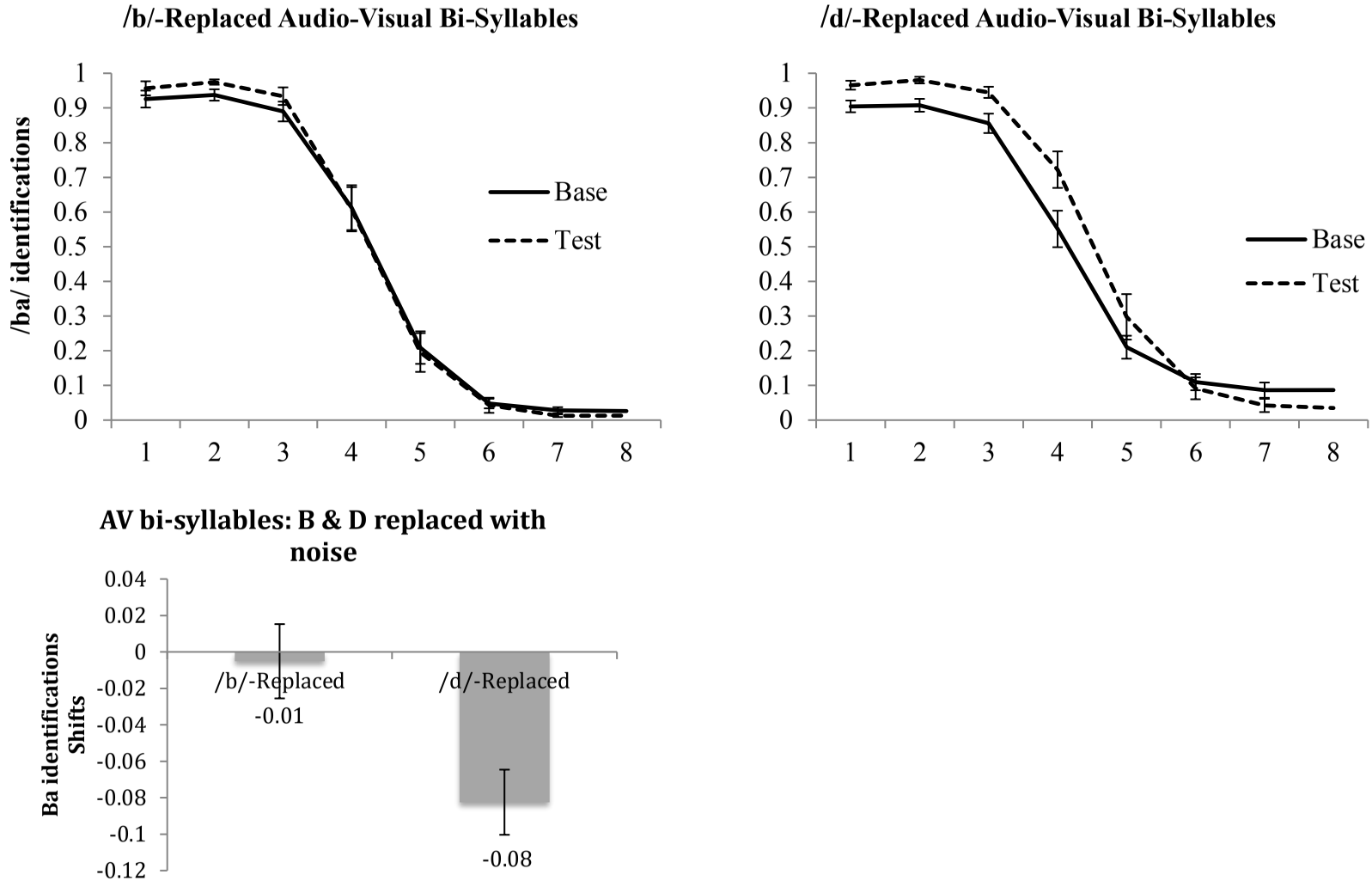
Figures 2.2a and 2.2b depict the proportion of participant “ba” identifications for each continuum item at baseline (before adaptation) and test (post adaptation). Figure 2.2a displays data for subjects who received clear /**b**/ words during adaptation, while Figure 2.2b displays data for subjects who received clear /**d**/ words during adaptation. Figure 2.2c displays the identification shifts (baseline – test) for subjects of both conditions; identification shifts are averaged across the middle four continuum items for each condition.

Figure 2.3


/b/- Noise Replaced Words
/ba/ identifications
Base
Test
/d/- Noise Replaced Words
Base
Test
Words With Noise
Ba identification Shifts
/b/-Replaced
/d/-Replaced

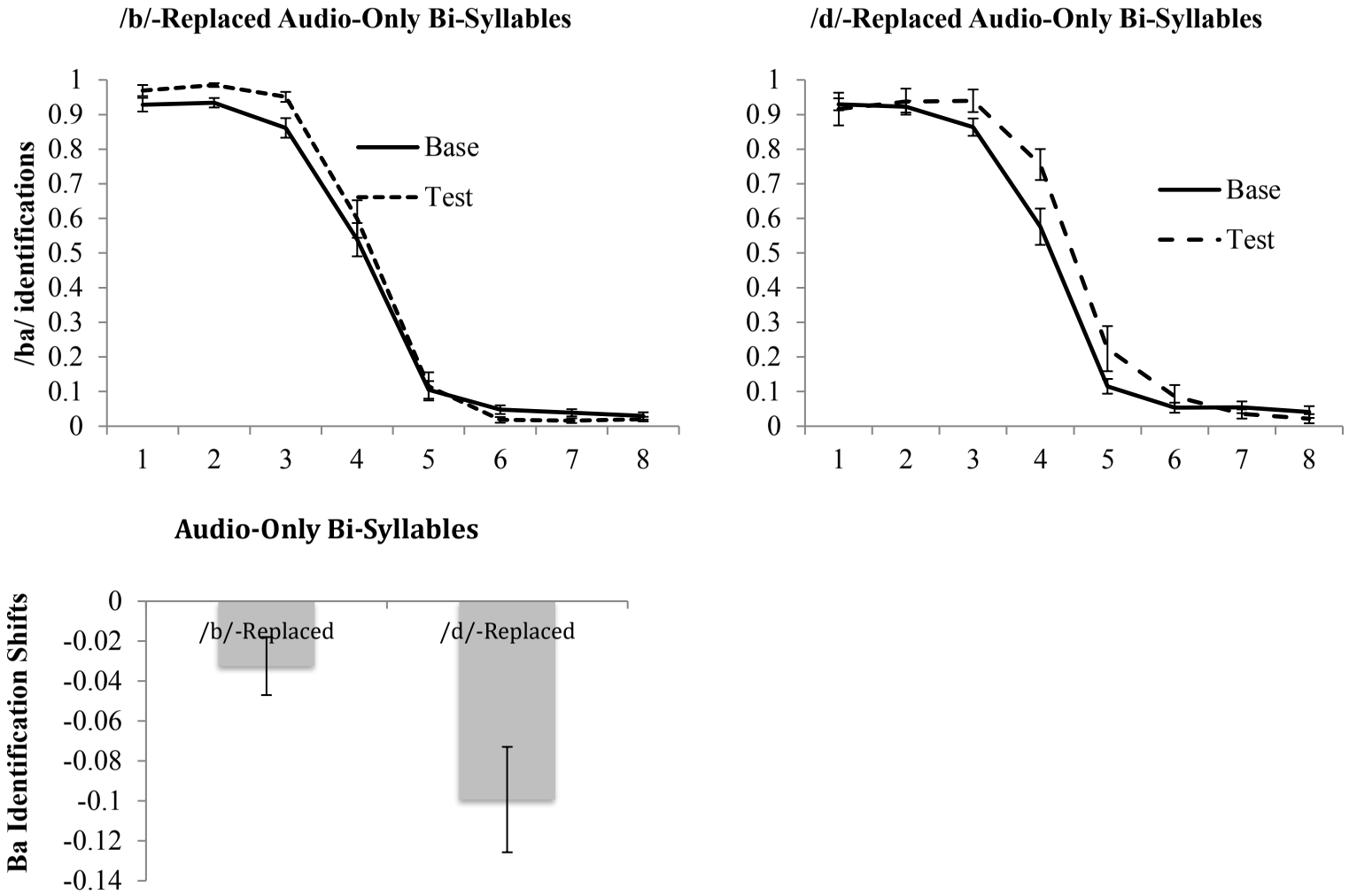
Figures 2.3a and 2.3b depict the proportion of participant “ba” identifications for each continuum item at baseline (before adaptation) and test (post adaptation). Figure 2.3a displays data for subjects who received words with /**b**/ replaced by signal-correlated-noise during adaptation, while Figure 2.3b displays data for subjects who received words with /**d**/ replaced by signal-correlated-noise during adaptation. Figure 2.3c displays the identification shifts (baseline – test) for subjects of both conditions; identification shifts are averaged across the middle four continuum items for each condition.

Figure 2.4


/b/-Replaced Audio-Visual Bi-Syllables
/ba/ identifications
Base
Test
/d/-Replaced Audio-Visual Bi-Syllables
Base
Test
AV bi-syllables: B & D replaced with noise
Ba identifications Shifts
/b/-Replaced
-0.01
/d/-Replaced
-0.08

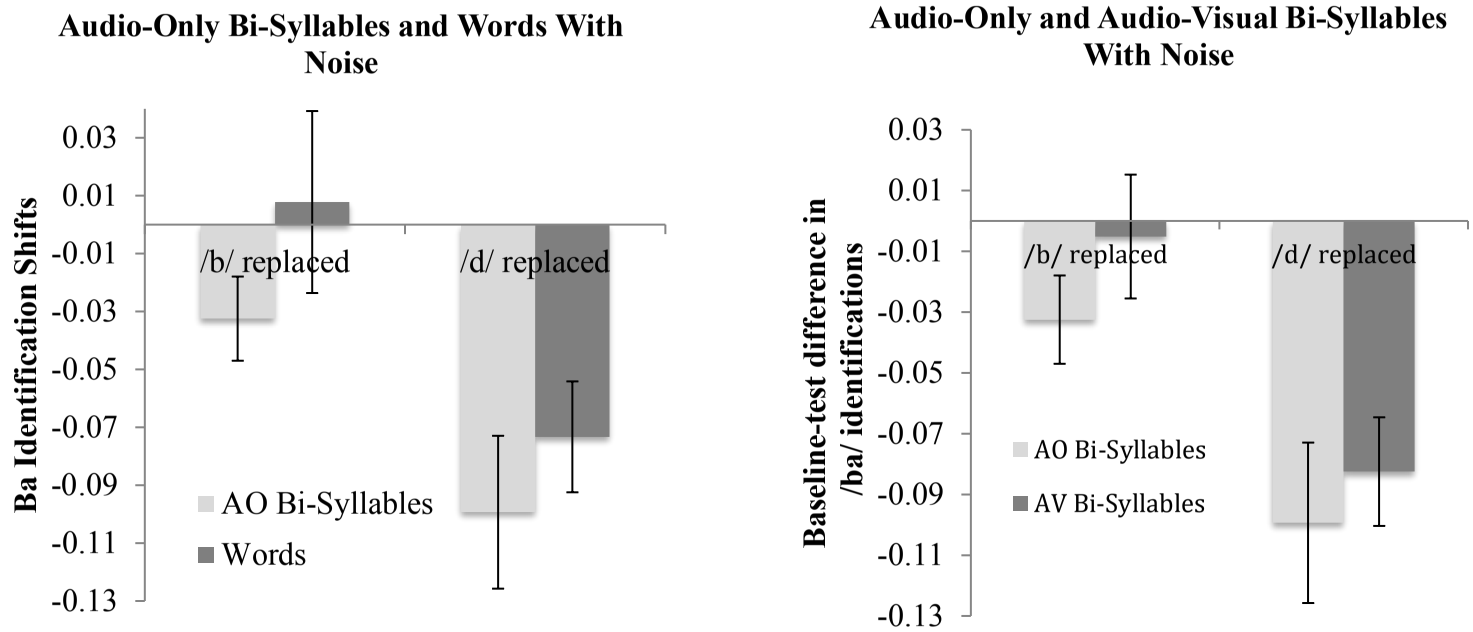
Figures 2.4a and 2.4b depict the proportion of participant “ba” identifications for each continuum item at baseline (before adaptation) and test (post adaptation). Figure 2.4a displays data for subjects who received audio-visual bi-syllables with /**b**/ segments replaced by signal-correlated-noise during adaptation, while Figure 2.4b displays data for subjects who received audio-visual bi-syllables with /**d**/ segments replaced by signal-correlated-noise during adaptation. Figure 2.4c displays the identification shifts (baseline – test) for subjects of both conditions; identification shifts are averaged across the middle four continuum items for each condition.

Figure 2.5


/b/-Replaced Audio-Only Bi-Syllables
/ba/ identifications
Base
Test
/d/-Replaced Audio-Only Bi-Syllables
Base
Test
Audio-Only Bi-Syllables
Ba Identification Shifts
/b/-Replaced
/d/-Replaced

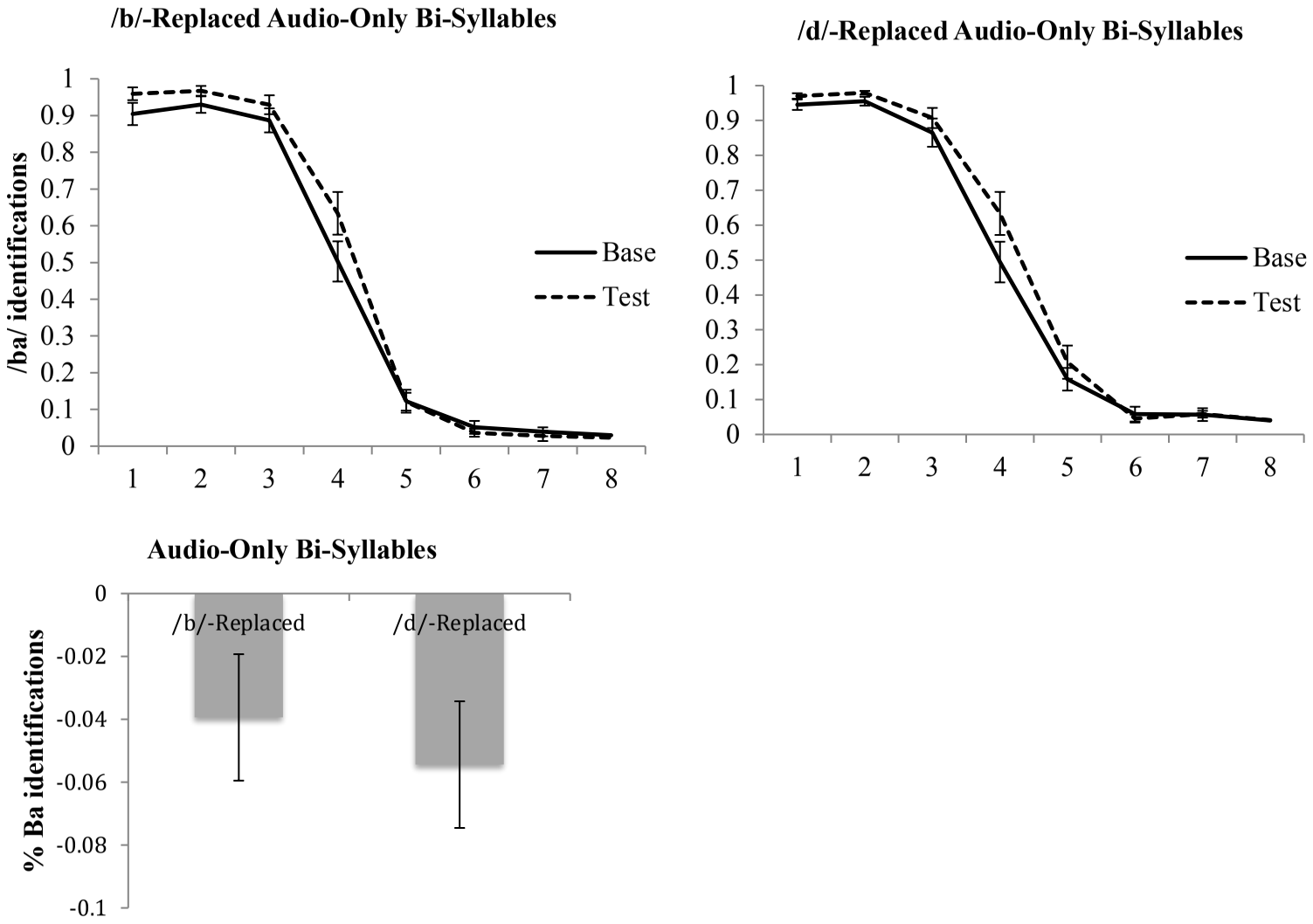
Figures 2.5a and 2.5b depict the proportion of participant “ba” identifications for each continuum item at baseline (before adaptation) and test (post adaptation). Figure 2.5a displays data for subjects who received audio-only bi-syllables with /**b**/ segments replaced by signal-correlated-noise during adaptation, while Figure 2.5b displays data for subjects who received audio-only bi-syllables with /**d**/ segments replaced by signal-correlated-noise during adaptation. Figure 2.5c displays the identification shifts (baseline – test) for subjects of both conditions; identification shifts are averaged across the middle four continuum items for each condition.

Figure 2.6



Figures 2.6a depicts the shift proportion of "ba" identifications from baseline (before adaptation) and test (after adaptation) for /b/ and /d/ noise replaced conditions in the audio-only bi-syllable and lexical (words with noise) context conditions; identification shifts are averaged across the middle four continuum items for each condition.
Figures 2.6b depicts the shift proportion of "ba" identifications from baseline (before adaptation) and test (after adaptation) for /b/ and /d/ noise replaced conditions in the audio-only bi-syllable and multisensory (audio-visual bi-syllables) context conditions; identification shifts are averaged across the middle four continuum items for each condition.

Figure 2.7


/b/-Replaced Audio-Only Bi-Syllables
/ba/ identifications
Base
Test
/d/-Replaced Audio-Only Bi-Syllables
Base
Test
Audio-Only Bi-Syllables
% Ba identifications
/b/-Replaced
/d/-Replaced

Figures 2.7a and 2.7b depict the proportion of participant “ba” identifications for each continuum item at baseline (before adaptation) and test (post adaptation). Figure 2.7a displays data for subjects who received audio-only bi-syllables with /**b**/ segments replaced by fixed amplitude noise during adaptation, while Figure 2.7b displays data for subjects who received audio-only bi-syllables with /**d**/ segments replaced by fixed amplitude noise during adaptation. Figure 2.7c displays the identification shifts (baseline – test) for subjects of both conditions; identification shifts are averaged across the middle four continuum items for each condition.

Figure 2.8

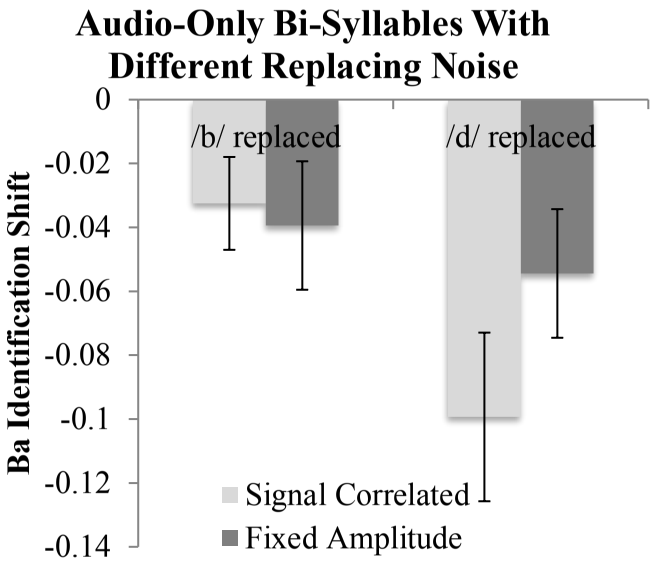
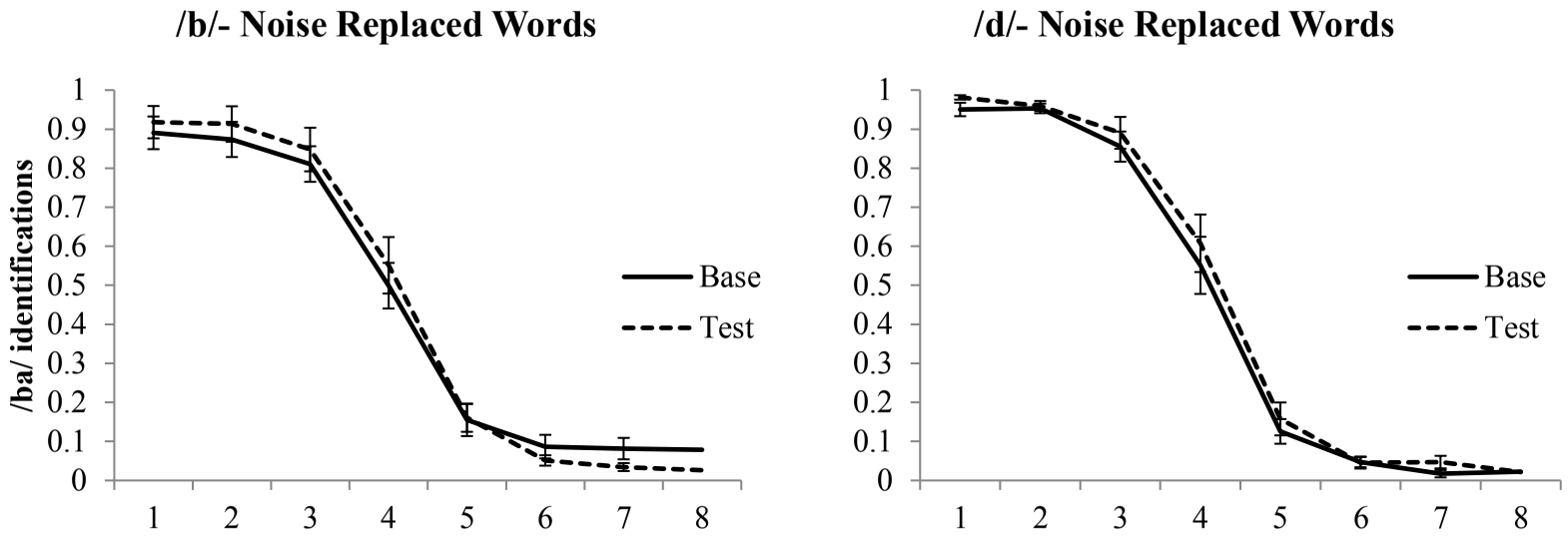
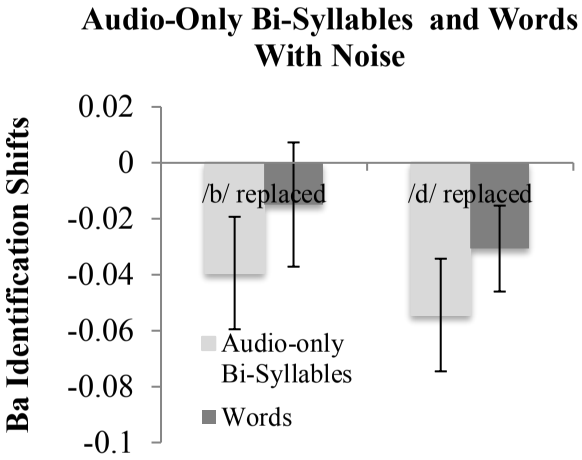


Figure 2.8 shows the shift in "ba" identifications between baseline (before adaptation) and test (after adaptation) for /b/ and /d/ contexts replaced by signal correlated and fixed amplitude noise; identification shifts are averaged across the middle four continuum items for each condition.

Figure 2.9

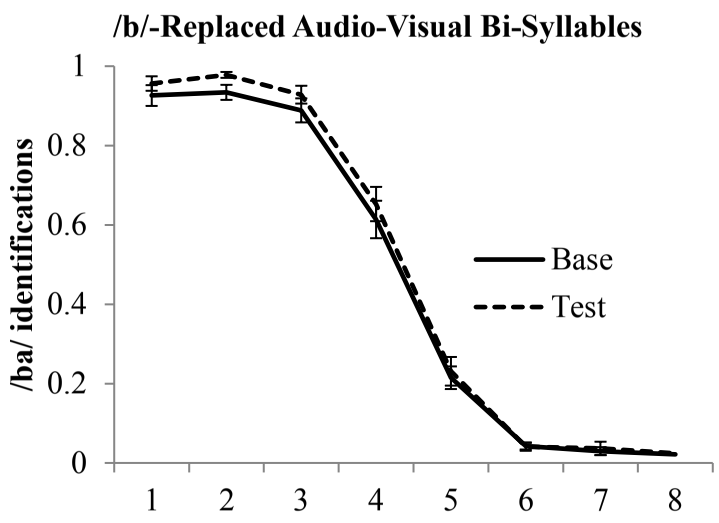

/b/- Noise Replaced Words
/ba/ identifications
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1 2 3 4 5 6 7 8
Base
Test

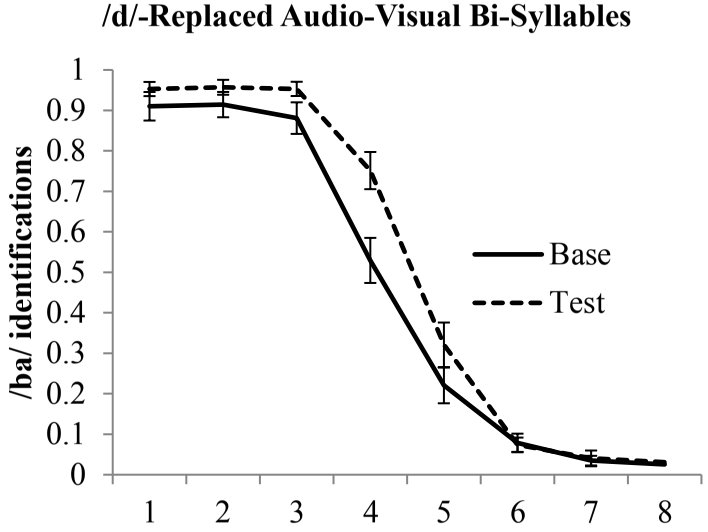
/d/- Noise Replaced Words
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1 2 3 4 5 6 7 8
Base
Test

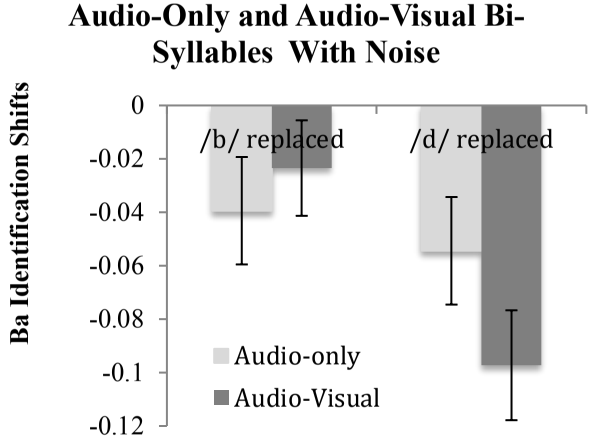

Audio-Only Bi-Syllables and Words With Noise
Ba Identification Shifts
0.02
0
-0.02
-0.04
-0.06
-0.08
-0.1
/b/ replaced
/d/ replaced
Audio-only Bi-Syllables
Words

Figures 2.9a and 2.9b depict the proportion of participant “ba” identifications for each continuum item at baseline (before adaptation) and test (post adaptation). Figure 2.9a displays data for subjects who received audio-only words with /**b**/ segments replaced by fixed amplitude noise during adaptation, while Figure 2.9b displays data for subjects who received audio-only words with /**d**/ segments replaced by fixed amplitude noise during adaptation. Figure 2.9c displays the identification shifts (baseline – test) for subjects of both conditions relative to corresponding audio-only bi-syllable conditions; identification shifts are averaged across the middle four continuum items for each condition.

Figure 2.10


/b/-Replaced Audio-Visual Bi-Syllables
/ba/ identifications
1
0.8
0.6
0.4
0.2
0
1 2 3 4 5 6 7 8
Base
Test


/d/-Replaced Audio-Visual Bi-Syllables
/ba/ identifications
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1 2 3 4 5 6 7 8
Base
Test


Audio-Only and Audio-Visual Bi-Syllables With Noise
Ba Identification Shifts
0
-0.02
-0.04
-0.06
-0.08
-0.1
-0.12
/b/ replaced
/d/ replaced
Audio-only
Audio-Visual

Figures 2.10a and 2.10b depict the proportion of participant “ba” identifications for each continuum item at baseline (before adaptation) and test (post adaptation). Figure 2.10a displays data for subjects who received audio-visual bi-syllables with /**b**/ segments replaced by fixed amplitude noise during adaptation, while Figure 2.10b displays data for subjects who received audio-visual bi-syllables with /**d**/ segments replaced by fixed amplitude noise during adaptation. Figure 2.10c displays the identification shifts (baseline – test) for subjects of both conditions relative to corresponding audio-only bi-syllable conditions; identification shifts are averaged across the middle four continuum items for each condition.

Figure 2.11

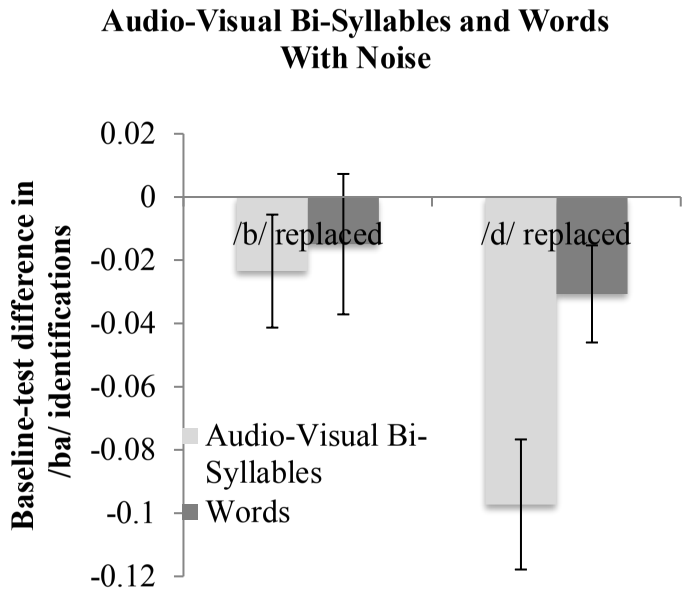
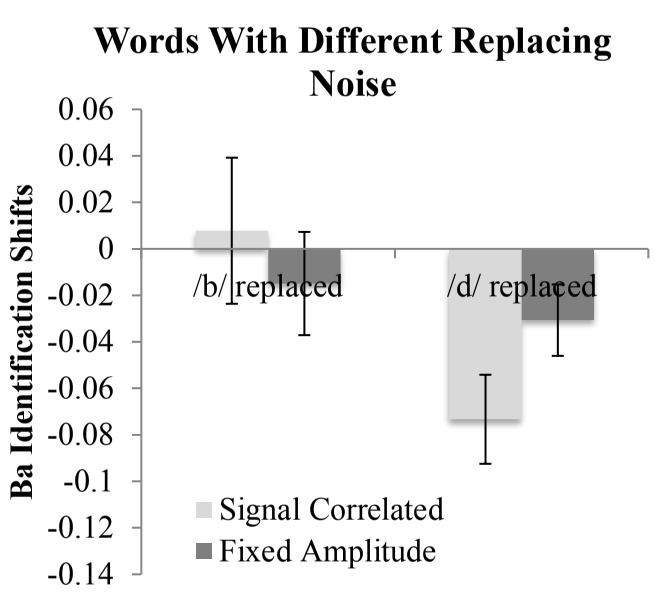
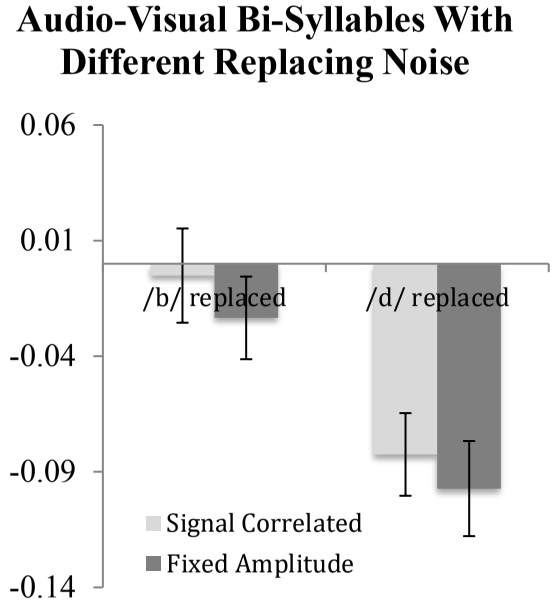


Figure 2.11 shows the shift in "ba" identifications between baseline (before adaptation) and test (after adaptation) for the /b/ and /d/ replaced by fixed amplitude noise conditions across the multisensory (audio-visual bi-syllables) and lexical (audio-only words) contexts. All values were averaged across the middle four continuum items.

Figure 2.12





Figures 2.12a and 2.12 b show the shift in "ba" identifications between baseline (before adaptation) and test (after adaptation) for the /b/ and /d/ replaced by fixed amplitude noise and signal-correlated-noise. Figure 2.12a displays means from the lexical context conditions (words with replacing noise) while 2.12b displays the means from the multisensory context conditions (audio-visual bi-syllables with replacing noise). All values were averaged across the middle four continuum items.

Chapter 3

Do Lexical Influences on Speech Perception Take Place Before, During, and/or After Multisensory Integration?

Do Lexical Influences on Speech Perception Take Place Before, During, and/or After Multisensory Integration?

When speaking to someone in a noisy environment, multiple sources of information can support the perception of speech. One is lexical information: the information provided by the word context the speech signal occurs in. The second is multisensory information, which includes (but is not limited to) the speaker's visible articulations. While both lexical and multisensory information are known to support speech perception, few investigations have examined speech perception in contexts that contain both sources. This gap raises questions concerning the processes by which the mechanism for speech perception combines these two sources of information.

A review of the speech literature reveals that lexical and multisensory information seem to support speech perception in similar ways. For example, both lexical and multisensory information can improve the perception of speech in noise (Miller, Heise, & Lichten, 1951; Hirsh, Reynolds, & Joseph, 1954; Sumby & Pollack, 1954; Grant & Seitz, 2000). Similarly, both can improve the perception of acoustically degraded speech, such as noise-vocoded speech (e.g. Bernstein et al., 2014; Davis et al., 2005) and both can bias the perception of phonetically ambiguous speech segments (e.g. the "Ganong effect"; Ganong, 1980; Bertelson, Vroomen, & De Gelder, 2003). But do these similar findings imply a similar process of incorporating lexical and multisensory information into the speech percept?

The seminal work of Brancazio (2004) has addressed this question. This study built on a multisensory illusion known as the McGurk effect: the finding that visual

speech can change how discrepant auditory speech is heard (e.g. audio ‘ba’ + visual ‘da’ is often heard as ‘da’; audio ‘ba’ + visual ‘ga’ is often heard as ‘ga’ or ‘da’; McGurk & MacDonald, 1976; MacDonald & McGurk, 1978). Brancazio (2004) measured the occurrence of visually dominated McGurk illusions produced by all combinations of audio-word or nonword + visual-word or nonword (e.g. audio nonword ‘beaf’ + visual word ‘deaf’; see also Sams et al., 1998; Barutchu et al., 2008). Brancazio (2004) found more McGurk illusions for audio-nonword + visual-word (e.g. audio ‘beaf’ + visual ‘deaf’) combinations than for audio-word + visual-word combinations (e.g. audio ‘band’ + visual ‘dand’), and that overall, McGurk effects were more common when they formed words.

Brancazio (2004) discusses his results in reference to a framework of the McGurk effect in which multisensory information is first integrated and then assigned to a phonetic category (See Figure 3.1). This model is well illustrated by a subsequent study conducted by Brancazio and Miller (2005). This study built on prior work by Green and Miller (1985), who found that presenting a fast visual ‘pi’ with items from an auditory ‘bi’-‘pi’ (which varied in voice-onset-time) continuum resulted in more items from that continuum being identified as ‘pi.’ This finding indicated that participants had integrated visual speaking rate information with auditory voice-onset-time. Brancazio and Miller (2005) built on this finding by combining items from an audio-only ‘bi’-‘pi’ continuum with fast and slow visual-only ‘ti’.

In the context of auditory ‘bi’ (or auditory ‘pi’) + visual ‘ti’, the McGurk effect would be characterized by reports of *hearing* ‘ti’ (or ‘di’) whereas “McGurk failures”

would be reports of ‘bi’ or ‘pi.’ Brancazio and Miller (2005) found that even when participants reported hearing ‘bi’ or ‘pi’ (McGurk failures) the speaking rate information from the visual stimulus influenced the phonetic boundary between ‘bi’ and ‘pi’. This meant that there were more ‘pi’ responses with fast visual stimuli, indicating that visual rate information had been integrated with the auditory voice-onset-time information, despite the participant’s response still being consistent with the auditory stimulus.

Consistent with the framework first introduced by Brancazio (2004), Brancazio and Miller (2005) explain this result by suggesting that there are two stages of perceptual processing that influence the McGurk effect. In the first stage, multisensory information is combined. In the second stage, that merged sensory information is evaluated and assigned to the phonetic category it most closely resembles. If this category happens to be the same category as the auditory signal, the McGurk effect will fail, but this does not mean that multisensory integration did not occur. Instead, this indicates that the multisensory integration process yielded an output that was closer to the phonetic category of the auditory stimulus than it was to other phonetic categories.

This interpretation bears on the lexical influences observed by Brancazio (2004). For example, consider a hypothetical comparison of McGurk rates produced by the audio-word + visual-nonword combination of audio ‘band’ + visual ‘dand’ and the audio-nonword + visual-word combination of audio ‘beaf’ + visual ‘deaf.’ In a model that assumes lexical information influences the McGurk effect responses *post* multisensory integration—during the *phonetic categorization* stage—then the combination of ‘beaf’ and ‘deaf’ might result in an integrated output that is 50% ‘d’. If so, then the phonetic

categorization process could be influenced by the lexical information that 'deaf' but not 'beaf' is a word and thus produces McGurk effects more than 50% (i.e. 'deaf' identifications > 50%). Similarly, while 'band' + visual 'dand' might also produce an integrated output approximating 50% 'd', the knowledge that 'band' but not 'dand' is a word could suppress 'd' categorizations (i.e. 'dand' identifications < 50%; fewer McGurk effects).

Next consider an account proposing lexical influences *during* integration. Under such an account, the lexical information for 'band' or 'deaf' could skew the integrated output to be more or less 'd' like. Under such an account, the fact that the integrated output was skewed away from 50% would bias the subsequent categorization and thus produce the same bias in identification described for a lexical processing during categorization account. Thus, while intriguing, the results reported by Brancazio (2004) were inconclusive with respect to the question of what part of the process during audio-visual speech perception is sensitive to lexical information.

A more theoretically perplexing result from Brancazio (2004) is that audio-word + visual-word (e.g. audio 'belt' + visual 'dealt') combinations produced fewer McGurk effects ('dealt' responses) than did audio-nonword + visual-word (audio 'beaf' + visual 'deaf'). This result suggests that at whichever stage of multisensory speech identification the lexical information was processed, it had access to the lexical information from the individual auditory and visual streams. This finding could indicate that lexical information is assessed early in speech processing, potentially prior to or during integration. Alternatively, this result could be explained by a post integration lexical

process that considers words with a certain degree of phonetic similarity to the integrated output.

Part of the interpretative difficulties with these results is that they measure the lexical effects on the McGurk effect in very general terms. Lexical information was either present (words) or absent (nonwords), and the results of these conditions were then averaged across items with the same word/nonword audio-visual structure. Thus a great deal of item level variation is lost, and this variation can be large (e.g. 4.04% [audio 'bay' + visual 'gay'] – 50.57% [audio 'bod' + visual 'god']; audio 'bay' + visual 'gay' sometimes perceived 'bay' or 'gay' but also sometimes as 'they' or 'lay'). Understanding the relationship between individual McGurk items and the McGurk effects they produce will provide valuable information for evaluating the influence of lexical information in audio-visual speech identification.

To better understand how lexical information interacts with multisensory information during speech perception, this project uses a quantitative measure of lexical information: lexical frequency. Lexical frequency is already known to influence speech perception, as is shown by lexical decision latencies (Marselen-Wilson, 1987) being shorter, and speech in noise recognition (Pollack, Rubenstein, & Decker, 1960) being improved, for higher frequency words. Here we investigate if lexical frequency can predict the McGurk effect.

Motivation for Analyses

To evaluate the lexical influences on the McGurk effect we will run two types of analyses. In the first set, we test the relationship between auditory and visual word lexical

frequency on auditory and visual word identifications. As both the auditory and visual words are present in the sensory input, if integration processes lexical information it should process the information about both auditory and visual words. Thus, if lexical information influences *multisensory integration*, it should be most apparent in the interaction between auditory or visual word lexical frequency. For example, visual word identifications should be most common for high frequency visual words combined with low frequency auditory words.

In the second series of analyses we examine the relationship between fusion word identifications (see below for details) and the corresponding lexical frequency of those words. As fusion words are not part of the sensory input to integration, processing of their lexical frequency should occur after integration. Thus, if lexical information can influence speech perception *post integration*, during the categorization stage, then there should be a strong relationship between fusion word frequency and fusion word identifications.

Finally, it is important to note that lexical processing during integration and lexical processing during categorization are not mutually exclusive hypotheses. It can certainly be the case that we will find evidence in support of both accounts. This would be indicated by significant effects in both classes of tests.

Experiment 1

Experiment 1 establishes item level differences in the identification of audio-word + visual-word McGurk stimuli. The basis of these differences comes in part from the different identifications associated with individual McGurk items, as well as the lexical

frequency of the words that those items are composed of. The McGurk effect can manifest as either a visually driven auditory percept in which participants report hearing the visual signal (e.g. audio 'bore' + visual 'gore' is heard as 'gore'; what we term 'McGurk-visual' responses) or percepts in which participants hear a speech sound that is present in neither the audio or visual signals (e.g. audio 'bore' + visual 'gore' is heard as 'door'; what we term 'McGurk-*fusion*' responses; see Alsius et al., 2018). Half of the McGurk items in Experiment 1 were expected to produce visual dominance type effects while the other half were expected to produce fusion type effects.

Method

Participants

Participants were 20 native English speakers (14 female) from the University of California, Riverside. All participants had normal hearing and normal or corrected to normal vision. All participants were compensated with course credit.

Materials

All stimuli were produced in a single recording session by a male, native monolingual English speaker. The speaker had lived in southern California for approximately 4 years prior to recording. He was digitally audio-video recorded uttering each of our word items at 30 frames-per-second (fps) at a size of 640 x 480 pixels. The items were 60 minimal word pairs with the audio-visual components differing only in the initial consonant which could be either b/v (e.g. audio 'boat' + visual 'vote'), b/d (e.g. audio 'bait' + visual 'date'), b/g (e.g. audio 'buy' + visual 'guy'), p/c (e.g. audio 'pod' + visual 'cod'), p/t (e.g. audio 'poll' + visual 'toll'), or m/n (e.g. audio 'might' + visual

'night'). The first author selected an audio-visual temporal alignment that both appeared to be synchronous and changed the percept of the auditory channel (i.e. produced the McGurk effect). The final stimulus showed the talker's entire face from the crown of his head to his shoulder.

The primary criteria for items used in this experiment were minimal word pairs that would support the McGurk effect. After selecting our word items, we tabulated the lexical frequency for each auditory and visual word based on the log word frequencies reported by Brysbaert & New (2009). During data analysis we identified fusion words that resulted from our stimuli, and we similarly tabulated the lexical frequency for these words. All auditory stimuli were presented through sound insulated headphones at an average of 70db.

Procedure

Participants were presented with all 60 McGurk items as well as with the 60 audio-alone items corresponding to the McGurk stimulus auditory channel. Items were blocked by audio-visual vs. audio-alone stimulus type. Participants were instructed to attend to each utterance and to use the keyboard to type the word they *heard* the talker say. Participants were allowed to view their responses as they typed them and were instructed to correct any errors before proceeding to the next trial. Each audio-visual trial included a fixation point at the location of the talker's lips that was present for 600ms immediately preceding the appearance of the talker's face. Participants received each stimulus three times during the experiment and stimulus presentation was randomized

within the audio-only and audio-visual blocks (Barutchu et al., 2008). Block order was randomized across subjects. The total experiment lasted about 20 minutes.

Results

Analysis of the Experiment 1 results began by designating three categories of participant responses: identifications consistent with the auditory stimulus (McGurk-failures), identifications consistent with the visual stimulus (McGurk-visuals), and identifications that were not consistent with either the auditory or visual stimulus (McGurk-fusions). Overall the stimuli produced reliable McGurk effects. Pooling across all sixty items, only 27.9% of responses were McGurk-failures, and 49.4% of responses were McGurk-visual responses. A complete summary of the identification scores for the McGurk items and their corresponding audio-only stimuli is provided in Table 3.1.

Lexical Influences on the McGurk Effect During Integration

Lexical information is present in the auditory and visual component stimuli of the McGurk tokens. Accordingly, if lexical information influences multisensory integration, then there should be a correlation between word identification and the lexical frequency of the auditory and visual components of the stimuli.

We examined this possibility by first testing a pair of item analyses that correlated auditory or visual word identification rates (excluding fusion identifications) with the corresponding word lexical frequency. The correlation between auditory word identifications and auditory word lexical frequency was (r [59] = .105, p = 0.421). The relationship between visual word identifications and the lexical frequency of visual words was (r [59]= -0.207, p = .112). While these are null effects, and thus are difficult to

interpret, they may indicate that lexical frequency does not bear on the components before or during the integration process, as such.

While these correlations are helpful in characterizing the overall patterns in the data, they may not be the best analysis for testing the predictions of integration and categorization accounts of lexical processing. This is because by averaging across participants to form item level effects, we lose a great deal of subject dependent variability, and there is evidence of substantial individual differences in the McGurk effect (Strand, Cooperman, Rowe, & Simenstad, 2014; Ujiie, Asai, & Wakabayashi, 2018). Furthermore, these correlations lack information about how this lexical information might *interact* during the perceptual process. If lexical information is processed during integration, then there may be an *interaction between auditory and visual lexical frequency* as both are present in the sensory input to integration.

Accordingly, we next conducted a series of linear mixed effect analyses. These analyses used subject and McGurk item as random intercepts and word identifications as the outcome. We first replicated the results from our correlations: auditory word identifications were not predicted by auditory word frequency ($\beta = 0.027$, $SE = .0331$, $t = 0.82$, $p = .416$) nor were visual word identifications predicted by visual word frequency ($\beta = 0.061$, $SE = .039$, $t = -1.561$, $p = .124$).

Finally we ran an analysis with fixed effects for lexical frequency of the auditory and visual words *and their interaction* in predicting auditory word identifications. The results of this analysis support the audio-visual interaction hypothesis, the model produced a significant interaction between auditory and visual lexical frequency ($\beta =$

0.067, *SE* = .029, *t* = 2.284, *p* = .026; see also Table 3.2 for the complete results; see also Figure 3.2). This finding is consistent with lexical influences during integration.

Lexical Influences on the McGurk Effect During Post-Integration Categorization

As the sensory input does not include fusion words, if we find a correlation between word identification and fusion word lexical frequency, then there is a post integration lexical process. As with the test of the integration hypothesis, we began with an item analysis that correlated word identification rates with their corresponding word lexical frequency, this time focusing on fusion word identifications and their frequency.

There was some variability in participant identifications of the McGurk items, so not all McGurk-fusion responses corresponded to the same identification, even for a single item. For this reason, in calculating the McGurk-fusion rate, we only included responses that a) formed an English word, and b) when there were multiple responses across subjects that formed words, we only counted the most common response as the "fusion" response (See Table 3.1). These criteria for McGurk-fusions resulted in 9% of non-audio/non-video responses being rejected. When combined with auditory and visual word identifications, our included fusions accounted for 92% of responses.

We found a significant correlation between fusion word identifications and fusion word lexical frequency ($r[38] = .523$, $p < .001$). To control for subject effects we ran a linear mixed effect analysis with fusion word identifications as an outcome and item and subject as random intercepts, and still found fusion word lexical frequency to be a

significant fixed effect (β = 0.089, *SE* = .029, *t* = 3.052, *p* = .004). These results are consistent with a lexical processing post integration account.

Fusion Word Frequency Effects on the McGurk Effect During Integration

The above tests assume that lexical influences of fusion words must follow integration. It is however conceivable that fusion words could influence integration, perhaps through feedback from a post integration process or perhaps due to the phonetic similarity of fusion words to the auditory and visual words. To evaluate these possibilities, we next ran a series of linear mixed effects analyses that all included random intercepts for subject and item, while testing for interactions between auditory, visual, and fusion word frequency in predicting auditory, visual or fusion word identifications. *If fusion word frequency affects word identification during integration, then there should be an interaction between those word frequencies and the auditory and visual word frequencies*.

We failed to find support for this hypothesis. No test showed the predicted interaction between auditory, visual, and fusion word frequency[8] (See Table 3.3 for a complete summary of results). We find no evidence that the lexical information of fusion words influences word identification through effects on integration. This challenges accounts that assume that lexical processing might feedback to multisensory integration.

Discussion

These results build on previous research (e.g. Brancazio, 2004; Baruchtu et al., 2008) further quantifying and characterizing the effect of lexical information on the

[8] However, the analysis of visual word identifications did return a significant interaction between auditory and visual word frequency (β = -0.35, *SE* = .155, *t* = -2.259, *p* = .03)

McGurk effect. Where prior studies were able to establish a word percept bias for McGurk stimuli, these results show that this bias is, to some extent predictable, by the frequency of that word in the listener's language.

Our interpretation of these data is influenced by the work of Brancazio (2004) who intimated that the McGurk effect is dependent on two separate processes: multisensory integration and phonetic categorization (See Figure 3.1). It seems that the relationship between lexical frequency and word identification is most direct among fusion McGurk percepts. This finding is consistent with a model in which lexical information is assessed *after* audio-visual integration. Fusion words can only exist as a product of integration. Because they do not exist in the sensory input as such, the effects of fusion words must occur after integration.

Lexical processing during integration should include lexical information from the sensory input, that is, from the auditory and visual words. By failing to find interactions between fusion and auditory and visual lexical frequency, we fail to find support for any sort of feedback from later processes to integration.

However, we also find some evidence for lexical processing *during* integration. We found that the lexical frequency of auditory and visual words interacted with each other, but not with fusion word lexical frequency. Any account that proposes lexical processing during integration should assume that both auditory and visual lexical information are processed in parallel as both of these streams make up the sensory input to integration. Thus our finding that both auditory and visual word identifications are best predicted by the interaction between auditory and visual lexical frequency is consistent

with a lexical processing during integration account. It seems we have evidence in support of lexical processing during *both* integration *and* categorization.

A potential limitation of the current experiment comes from our decision to present our McGurk stimuli multiple times to each subject. Repetitive presentations are common within McGurk studies, including the lexical McGurk study by Barutchu et al., (2008). However, there is some research indicating that the effects of lexical frequency might diminish with repetition (See Colombo et al., 2006 for a discussion). This raises the question of whether the results of this experiment are compromised by our use of item repetition.

We argue that the results of the current experiment are still valid despite the risk of repetition effects. There is no research establishing that our specific task would be sensitive to repetition effects, and while repetition effects have been found across paradigms, the size of those effects *is* task specific (e.g. Schilling, Rayner, & Chumbley, 1998). Furthermore, while repetition can reduce the difference between high and low frequency words, they generally do not eliminate those differences entirely (Forster & Davis, 1984). Importantly, it should be noted that repetition effects generally take the form of diminishing differences between low and high frequency words (e.g. Griffen & Bock, 1998; Scarborough, Cortese, & Scaroborough, 1977). In other words, the risk of repetition to our experiment would have been failing to find an effect of lexical frequency; that we did find effects of lexical frequency remains notable[9].

[9] Our motivation in Experiment 1 was to determine if lexical frequency affects the McGurk effect. That we found effects of lexical frequency on the McGurk effect means Experiment 1 was successful in this. However, there are competing theoretical predictions concerning how repetition might influence lexical

Experiment 2

One limitation of Experiment 1 is that it relied on word identifications of McGurk stimuli to infer about lexical processing at two different points preceding identification: integration and categorization. An important piece of information for discriminating when lexical influences occur is knowing the *phonetic identity* of the critical segments from the integration process, independent of lexical processing. Experiment 2 will offer a more focused investigation into this point.

Above we discussed two points during speech perception at which lexical information might be processed; either during the multisensory integration stage, or during the phonetic categorization stage of perception (See also Brancazio, 2004; Brancazio & Miller, 2005). However, we now consider a third point during the multisensory language process which lexical information could influence: before the speech signal even reaches the perceiver (see Figure 3.3).

The speech stimuli that perceivers process must first be spoken by a talker, and there is evidence that lexical information actually influences speech *production* (e.g. Balota & Chumbley, 1985; Pluymaekers, Ernestus, & Baayen, 2005; Jurafsky et al., 2001). For example, segments from high frequency words tend to be spoken faster than the same segments spoken as part of low frequency words (Pluymaekers et al., 2005). By shaping the stimuli that eventually get perceived, lexical information might influence the perceptual process before the perceiver even receives the sensory signal.

---

frequency effects in different tasks (see Colombo et al., 2006). As such, the subsequent experiments in this investigation will avoid repetition for the sake of interpretative ease.

The potential of these lexical effects on speech production to influence speech perception is important to the interpretation of Experiment 1. It is possible that the interaction between auditory and visual lexical information observed in Experiment 1 was not the result of lexical *processing during integration*, but instead the result of lexical influences on the production of the critical auditory and visual *segments* that were subsequently integrated. To investigate this possibility, Experiment 2 uses McGurk *syllables* that were excised from the McGurk words used in Experiment 1. If the perception of these syllables is found to be predictable by the lexical frequency of the words from which they were extracted, then some of the lexical effects on speech perception might be attributed to effects on production. *If lexical influences on speech production influence multisensory integration, then there should be an interaction between auditory and visual lexical frequencies on the perception of syllables extracted from McGurk words*.

A second feature of Experiment 2 concerns its dependent measure. Prior work has shown that even when a stimulus produces a robust McGurk effect, it is perceptually distinct from audio-visually congruent stimuli. For example, Rosenblum and Saldana (1992) found McGurk stimuli were readily discriminated from congruent audio-visual stimuli, and Brancazio (2004) found that McGurk stimuli influenced the rating of a percept's goodness. As Experiment 1 indicated that some of the variability in the McGurk effect could be accounted for by lexical information, we were also interested in whether lexical information could also account for these qualitative changes in the McGurk effect.

For this reason, in Experiment 2 we also measured goodness ratings associated with the McGurk stimuli.

Goodness ratings have been used to measure graded differences in speech perception across a number of paradigms (e.g. Miller & Volaitis, 1989; Allen & Miller, 2001; Evans, & Iverson, 2004; Drouin, Theodore, & Myers, 2016), including McGurk studies (Brancazio, Miller, & Pare, 2003; Brancazio, 2004). The method of measuring goodness ratings for McGurk stimuli involves asking participants to attend to a speech stimulus and report first a nominal percept (e.g. 'ba' vs. 'da' or 'bait' vs. 'date') followed by a numerical value indicating how good an example the stimulus was of the participant's ideal version of that item (e.g. Brancazio, 2004). While these measures are inherently subjective, there is evidence that they are generally consistent across participants (for auditory stimuli; Iverson & Kuhl, 1995), as well as correlate with acoustic parameters of stimuli and with identifications rates (Samuel & Kat, 1996; Allen & Miller, 2001; Brancazio et al., 2003). Experiment 2 will use goodness ratings to qualitatively evaluate the percepts associated with the McGurk syllables. If lexical information influences the perceptual quality of a McGurk effect, then goodness scores will correlate with lexical frequency.

In short, Experiment 2 will measure the identification and quality of McGurk syllables extracted from McGurk words. The use of these syllables extracted from words allows this experiment to test if lexical effects on speech *production* can affect speech identification.

Method

Participants

Participants were all native English speakers from the University of California, Riverside. Eleven participants (9 female) participated in a congruent control condition, and there were an additional 20 (12 female) who participated in the main McGurk experiment. All participants had normal hearing and normal or corrected to normal vision. All participants were compensated with course credit.

Materials

The stimuli of the current experiment were generated from the McGurk words of Experiment 1 by excising the initial vowel-consonant syllable from each item. Occasionally, in doing this, the McGurk effect ceased to occur. This posed a problem for the current experiment because we could not investigate the effects of lexical frequency on the McGurk effect if the stimuli were not reliably producing McGurk effects.

As such, for these stimuli, the temporal alignment of the auditory and visual signals in the syllables was adjusted until the researcher experienced McGurk effects. While these adjustments suited the needs of the current experiment, of facilitating the McGurk effect in syllables, they have the drawback that the results of this experiment cannot be compared directly to the results of Experiment 1. However, a comparison of word and corresponding syllable McGurk effects will be provided in Experiment 3[10].

[10] That this adjustment needed to be made, in of itself, is likely an indication of the nature of lexical processing as it suggests that the McGurk effect became frailer when isolated from its word context. Future work should to investigate this issue in more detail.

Though the current experiment uses syllables, not words, for the sake of caution we chose to eliminate items with auditory signals taken from the same word (see Colombo et al., 2006 and the above discussion). There were three pairs of McGurk items from Experiment 1 that met this criteria; audio 'beer' + visual 'veer'/ audio 'beer' + visual 'deer,' audio 'buy' + visual 'vie'/audio 'buy' + visual 'guy,' and audio 'bet' + visual 'vet'/ audio 'bet' + visual 'debt.' As the plurality of our McGurk stimuli used the format of audio b-word + visual v-word (to promote visual dominate responses), we removed these items from the listed pairs.

All stimulus editing was done in Final Cut Pro 5 software for Mac OSX. The congruent versions of each item were matched to the length of the shorter of the auditory and visual information in the corresponding McGurk items.

Procedure

In contrast to Experiment 1, to avoid repetition effects, each McGurk syllable was only presented to each participant once. Thus the McGurk group was presented with 57 McGurk syllables and the congruent group was presented the 114 congruent syllables that corresponded to the McGurk stimuli components.

For both groups, each trial of the experiment consisted of the participant being presented with an audio-visual stimulus and verbally reporting the initial syllable that they heard, followed by a numeric value ranging from 1 to 5 (Brancazio, 2004). Experiment instructions explained that a 5 was to indicate that the heard syllable was what the participant would consider an excellent example of that syllable, while a 1 would indicate a poor example of that syllable. During the experiment, the participants

gave their responses verbally, by speaking into a microphone. A researcher, stationed outside the sound booth listened to the participants' responses, and typed: 1) an orthographic transcription of the syllable the participant reported hearing, and 2) the numeric value that participant reported. The researchers were unaware of the predictions of the study. Furthermore, the researcher could only hear what the participant said, and not what stimulus was presented to the participant, and thus were blind to which condition each participant was assigned.

Each audio-visual trial included a fixation point at the location of the talker's lips that was present for 300ms immediately preceding the appearance of the talker's face.

Results

Stimulus Identifications

We began our analysis by confirming that the congruent syllables were reliably identified. We divided our congruent syllables into two groups, words that were congruent with the McGurk auditory stimulus (audio 'boat' + visual 'boat'; corresponding to the McGurk audio 'boat' + visual 'vote'), and those that were congruent with the McGurk visual stimulus (audio 'vote' + visual 'vote'). We tabulated the proportion of responses that each item was correctly identified and found reliable identifications for both groups (congruent tokens consistent with the McGurk auditory component: $M = .915$, $SE = .012$; congruent tokens consistent with the McGurk visual component: $M = .869$, $SE = .022$; see also Table 3.4).

We next examined the McGurk items, and found that in general, they produced reliable McGurk effects: reports of hearing the auditory syllable ($M = .198$, $SE = .023$)

was significantly smaller (t[56]= 5.67, p <.001, r = .604) than the rate of McGurk visual identifications (M = .513, SE = .039).

Lexical Effects of Auditory and Visual Words. We started with a dichotic outcome of McGurk failure (auditory identifications) and McGurk success (participant heard something other than the auditory stimulus; pooling McGurk fusion and McGurk visual responses) and using subject and McGurk item as random intercepts. We used the lexical frequency of the words from which the McGurk syllables were extracted, as our fixed effects. We tested our main hypothesis, *if the lexical influences on speech production influence the identification McGurk stimuli, then there will be an interaction between auditory and visual word frequencies in the McGurk effect for syllables*. We ran an analysis with fixed effects of the auditory and visual word lexical frequencies, and their interaction. We found significant effects of auditory lexical frequency (β = -2.010, SE = .591, z = -3.400, p = .007), visual lexical frequency (β = -1.838, SE = .607, z = -3.028, p = .003), and the interaction of these measures (β = .648, SE = .197, z = 3.288, p = .001[11]; see Table 3.5, see also Figure 3.4).

To better understand the nature of this effect, we next ran a simplified analysis that included only lexical frequency from the auditory and visual words, but excluded their interaction. This main effects test showed no effect of either lexical frequency. Thus it seems that the *interaction* between auditory and visual word lexical frequency is an important factor in predicting the McGurk effect in syllables extracted from those words.

[11] Though not reported (for the sake of brevity) it is worth noting that we also found an interaction between auditory and visual word frequencies in predicting visual identifications. This is another similarity between the results of this experiment and Experiment 1.

That the stimuli used in this experiment were only syllables extracted from words, it seems unlikely that this lexical effect on integration is related to lexical processing of the perceiver. It is more likely that these results reflect lexical effects on the *production of the stimuli* that the perceiver later integrated. These results support our hypothesis that lexical information can influence multisensory integration through the effect of lexical information on *speech production*. To our knowledge this is the first evidence of this kind of effect.

Lexical Influences on the Perceived Goodness of Syllables. In a final set of analyses we investigated if lexical information would influence subjects' goodness ratings. Using subject and item as random intercepts we failed to find an effect of lexical frequency on the ratings of the congruent syllables ($\beta = .031$, $SE = .048$, $t = .655$, $p = .513$). Nor did we find any effect of lexical frequency on the goodness rating in the McGurk syllables (see Table 3.6). However, when we included an interaction with the average goodness rating for the congruent syllables, we found several interactions with lexical frequency (See Table 3.6b). This indicates that the relationship between lexical frequency and McGurk syllable goodness is dependent on the goodness rating of the sensory streams that make up the McGurk stimulus.

Interestingly, these effects were absent when we analyzed the effect of auditory or visual lexical frequency alone (Tables 3.6c-d). This finding suggests that, as with the identifications, it is the *interaction* between auditory and visual word frequency that is important to predicting the goodness ratings of McGurk syllables. In short, it seems that

in addition to predicting syllable identification, the interaction between auditory visual lexical frequency can predict the perceived goodness of McGurk syllables.

Discussion

Perhaps the most important discovery of Experiment 2 is that multisensory integration reflects lexical influences on syllables extracted from words. It is worth noting that this could be driven by an effect of syllable frequency, which is potentially correlated with lexical frequency. Such an effect of syllable frequency would likely take the form of fewer McGurk effects (auditory identifications) for higher frequency auditory syllables. We are unaware of any research substantiating this relationship, however, it is an intriguing prospect that should be addressed in future work.

Another explanation for these results is that the McGurk effect is sensitive to the influences of lexical information on speech *production*. This finding is not only interesting in of itself, but also has implications for our interpretation of the results of Experiment 1. Recall that Experiment 1 revealed that the success of the McGurk effect in words was predicted by the interaction of the lexical frequency of auditory and visual words. In Experiment 1 this interaction could reflect the perceiver lexically processing the auditory and visual signals *during* multisensory integration. However, because we find the same interaction effect with syllables extracted from words another interpretation is possible. These results may indicate that at least some of the audio-visual lexical interaction of Experiment 1 is attributable to lexical influences on speech production structuring the auditory and visual sensory inputs to the integration process.

Experiment 3

To understand if the lexical influences on speech production observed in Experiment 2 can account for the entire audio-visual lexical interaction effect seen in the results of Experiment 1, Experiment 3 will measure perception of McGurk stimuli in two blocks of trials. The first block will be the non-lexical McGurk stimuli used in Experiment 2 consisting of syllables generated by excising the initial consonant-vowel from the McGurk words used in Experiment 1. The second block of trials will present lexical McGurk stimuli, consisting of the full words from which the syllables were extracted. In this way, the perception of items in the syllable block will inform us about lexical effects during speech *production* while the perception of the word items will inform us about lexical *processing* during perception. If there is lexical processing during speech perception, then there should be an interaction between syllable perception and lexical frequency. In this way, Experiment 3 provides a syllable-to-word McGurk comparison that was not possible previously due to the different stimulus editing procedures between Experiment 1 (words) and Experiment 2 (syllables).

One complication facing this experiment is that it requires the participants to be functionally exposed to each syllable stimulus twice; once as an isolated syllable and once as syllable within a word. These repetition effects can have theoretical implications (see Colombo et al., 2006 for a discussion). For example, within the TRACE framework (McClelland & Elman, 1986), these repetition effects would mean that prior exposure to a word would increase the resting activation of that lexical representation and, by virtue of feedback connections, the syllables associated with that lexical representation. That is,

hearing a syllable in the context of a word could change how that same isolated syllable is subsequently perceived (see McClelland, Mirman, & Holt, 2006 for a discussion). Thus, presenting our word block before our syllable block could result in substantial complications for interpretation. Therefore, Experiment 3 tested the syllable block first, followed by the word block.

The goal of Experiment 3 is to dissociate lexical influences on production from lexical processing during integration from lexical processing during phonetic categorization. We wanted a test of this question that would be more sensitive than simply comparing identification rates between syllable and word blocks. Two studies provide key insight into how we will conduct this test. First, Allen and Miller (2001) assessed how different types of contexts (i.e. speaking rate & lexical) influenced the goodness ratings of items from an auditory /p/ to /b/ continuum. Consistent with prior work, these researchers found that goodness ratings of the continuum items corresponded to each items' proximity to the phonetic boundary (proximity to the center of the continuum; see Figure 3.5a) with items on the phonetic boundary receiving the lowest goodness ratings. These researchers report an important finding: some contexts (e.g. lexical) were found to *primarily* influence the goodness rating of items that previously had low goodness ratings. These items were generally those along the phonetic boundary on the continuum (see Figure 3.5b) and the change in goodness rating corresponded with the contextually determined phonetic category. That is, the phonetic categorization process changed only the goodness ratings of continuum items around the phonetic

boundary. Unambiguous items further from the boundary were less prone to change from their initial goodness ratings.

This pattern contrasts with the effects of other types of contexts (e.g. speaking rate), which changed goodness ratings of *most* continuum items, irrespective of their proximity to the phonetic boundary (see Figure 3.5c). These latter contexts were assumed to affect feature integration, as opposed to simple phonetic categorization. Thus, the different patterns of goodness ratings were interpreted as reflecting different points of processing.

In a related study, Brancazio et al., (2003) also measured goodness ratings along an auditory /b/-/p/ continuum. These authors found that *visual* context also shifted the goodness ratings across most of the items on the continuum; a pattern very similar to what Allen and Miller (2001) attributed to effects on feature integration. On the basis of the results reported by Brancazio and his colleagues (2003), it seems that information processed by multisensory integration should produce the similar broad shifts in goodness ratings across items of varying distance from the phonetic boundary.

In short, these studies offer two findings relevant to the present investigation. First, they show that a percept's proximity to a phonetic boundary can be inferred by the goodness rating the perceiver assigns it— with lower goodness rated items likely being closer to the phonetic boundary. Second, these studies suggest that the integration and (post-integration) phonetic categorization processes produce two contrasting changes to goodness rating: 1) *integration* changes the goodness ratings of all (or most) items while

2) *phonetic categorization* only changes the goodness ratings of lowest rated items/items near the phonetic boundary (see Figure 3.5).

While the stimuli used in Experiment 3 will not use a speech continuum, the variability of the McGurk effect across items suggests that we will find a range of goodness scores for our stimuli and these goodness scores can be used to infer proximity to that boundary (Brancazio & Miller, 2005; Brancazio, 2004; Rosenblum & Saldana, 1992). Experiment 3 will assess how the lexical information provided by word contexts interacts with the goodness rating of the McGurk items. *If lexical processing influences multisensory integration itself, then the word contexts should change the goodness ratings across a range of items not restricted to items with syllables near the phonetic boundary* (e.g. Figure 3.5c). Alternatively, *if lexical processing influences the post-integration, phonetic categorization phase, then word context will only change the goodness ratings of the items that had the lowest rated syllables*[12] (e.g. Figure 3.5b).

Method

Participants

Twenty (15 female) native English speakers from the University of California, Riverside participated in Experiment 3. All participants reported having normal hearing and vision. All participants were compensated with course credit or $10.00 cash.

[12] These predictions are derived from the work of Allen & Miller (2001). As described in the main text, these authors found that two types of contextual information, auditory and lexical, produce two different patterns of goodness score change. Lexical context produced a change that was concentrated to the lower rated items while auditory context produced a change across the entire range of goodness scores. That both patterns, change of the lowest only vs. change across the entire range, were found demonstrates that the goodness scoring was capable of showing both patterns. Thus the determinant of which pattern of change occurred was the cognitive process associated with the different context, not an artifact of the goodness scores themselves.

Materials

The stimuli included the syllables that were used in Experiment 2, as well as the words from which those syllables were excised. Recall, that in making those syllables, temporal alignment adjustments were made, to ensure McGurk effects with the syllables. The McGurk word stimuli used in Experiment 3 were re-adjusted to retain the temporal adjustments made to the syllables (and thus, while derived from the same stimuli, are qualitatively different from the McGurk word items of Experiment 1). This will allow for a more direct comparison of McGurk effects between syllables and the words they were extracted from.

Procedure

The procedure for Experiment 3 was identical to the procedure for Experiment 2, the only difference being that Experiment 3 also included a block of 57 McGurk words (corresponding to the 57 McGurk syllables of the preceding block) which was presented immediately following the syllable block.

Results

McGurk Rates

We began our analysis by tabulating the proportion of subject responses that indicated auditory, visual, or fusion percepts for both McGurk syllable and word items (See Table 3.7). Within the syllables, there were robust McGurk effects, with significantly fewer auditory identifications (M = .191, SE = .023) than visual stimulus identifications (M = .474, SE = .033; t[50]=6.011, p <.001, r = .626). Identification of the

syllable that corresponded to the fusion words was rare (*M* = .117, *SE* = .020), and overall, occurred (infrequently) for only 27 of the 57 items.

Within the words, a similar pattern was found; visual identifications were the most common (*M* = .54, *SE* = .038), while auditory identifications were less common (*M* = .213, *SE* = .028). Also like the syllables, the visual identifications were significantly more common than the auditory (*t*[50]=5.703, *p* <.001, *r* = .606) identifications. Fusions were again the most rare (*M* = .113, *SE* = .022), and occurred for only 29 of the 57 items. It is interesting to note that relative to Experiment 1, there seems to be a shift towards increased visual identifications and decreased auditory identifications for the words, with no change in the rate of fusion identifications; this is likely attributed to the temporal alignment edits applied to the current stimuli.

Goodness Ratings

Auditory and visual identifications of the syllable stimuli corresponded to comparable goodness ratings (auditory: *M* = 2.707, *SE* = .126; visual: *M* = 2.727, *SE* = .096; see Table 3.7). Fusion type responses tended to have lower goodness ratings (*M* =1.437, *SE* = .151). Among the words, there was a general trend of improved goodness ratings with the visual identifications being the "most good" (*M* = 3.837, *SE* = .062), followed by the auditory (*M* = 3.545, *SE* = .118), and then the fusion (*M* = 2.002, *SE* = .136) identifications.

Lexical Influences on McGurk Effects in Syllables

We began by using McGurk success (vs. failure/auditory identification) for the syllable block as the outcome, and subject and item as the random intercepts. In two

separate analyses we replicate our results from Experiment 2, finding that the McGurk effect in syllables was predicted by the interaction between auditory and visual word frequencies (see Table 3.8), but not the lexical frequency of the fusion word (β= -0.014, *SE* = .149, *z* =-.094, *p* = .925).

Lexical Influences on McGurk Effects in Words

Again using subject and item as random intercepts, we found significant effects of auditory (β= 2.077, *SE* = .685, *z* = 3.032, *p* = .002) and visual (β= 2.153, *SE* = .709, *z* = 3.038, *p* = .002) word frequencies, as well as their interaction (β= -.744, *SE* = .229, *z* = -3.250, *p* = .001), for predicting the auditory identifications of McGurk words (i.e. "McGurk failures"). These results replicate our findings from Experiment 1[13]. This fact is notable as these stimuli were re-edited (audio-visually realigned) and therefore subtly different from the stimuli of Experiment 1. Importantly, in light of the similar findings from Experiment 2, it seems possible that this interaction between auditory and visual word frequency is not the result of lexical processing, but of lexical influences on production.

Does Speech Production Mediate the Word Identification Auditory and Visual Lexical Frequency Interaction?

Tests were conduct to examine if the current word identifications show an interaction between auditory and visual lexical frequency as a result of: 1) lexical

[13] We also investigated the effect of fusion word lexical frequency on the McGurk effect in words. We failed to find an effect of fusion word lexical frequency (β= .047, *SE* = .178, *z* = .265, *p* = .791). When we added McGurk rate from the syllables we found no interaction (β= -.067, *SE* = .189, *z* = -.352, *p* = .725). However, consistent with the implications of the above correlation analyses, and the results of Experiment 1, we did find that the lexical frequency of fusion words predicted the occurrence specifically of *fusion* McGurks in the words (β= 1.140, *SE* = .238, *z* = 4.783, *p* = .001).

influences on production (as was found in Experiment 2); or 2) influences during integration. For this purpose, we tested the effect of syllable McGurk rate on word McGurk rates. We first ran an analysis using McGurk success for syllables as a fixed effect, and found it to be a significant predictor of McGurk word success (β= 1.671, *SE* = .001, *z* = 1758.5, *p* < .001).

We next added fixed effects for the lexical frequency of auditory and visual words, as well as their interaction. From this analysis we found significant effects of lexical frequency (Audio: β= 1.813, *SE* = .754, *z* = 2.404, *p* = .016; Visual: β= 2.133, *SE* = .800, *z* = 2.667, *p* = .007; Interaction: β= -.744, *SE* = .259, *z* = -2.878, *p* = .004); but not for the syllable identification (β= 2.17, *SE* = 2.244, *z* =.967, *p* < .334; 3-way interaction: β=.206, *SE* = .252, *z* =.414, *p* = .414; both 2-way interactions with frequency had p >.4). This outcome suggests that lexical frequency may have an effect on word identification that is *not* accounted for by the integration of the syllables, as such.

There are two possible explanations for these results. First, this might indicate lexical processing during perception. This processing could reflect lexical effects during either the integration or phonetic categorization phases. Alternatively, this outcome could be another effect of speech production. We found evidence, both in the current experiment and Experiment 2, that lexical frequency effects on speech production influence speech perception in syllables. It is unlikely that these production effects are limited to influencing the perception of just the initial syllables. Whatever influence lexical frequency effects on production have on the perception of the word-initial syllables is also likely present in the segments following those initial syllables. In short,

this outcome poses many of the same questions that were raised during the analysis of Experiment 1.

Does Lexical Processing Take Place During Phonetic Categorization?

As stated, we are considering three points during speech perception that lexical information could influence identification: the production of speech stimuli, the integration of cross-sensory information, and categorization of the post-integrated output. Distinguishing which of these stages produces lexical effects could be illuminated by an analysis of goodness ratings (see above discussion).

We ran an analysis with word ratings being predicted by fixed effects for syllable goodness rating (1-5), syllable identification (auditory, visual, fusion word consistent, & "other"), word identification (auditory, visual, fusion word consistent, & "other"), and fusion lexical frequency. To control for the effects of lexical frequency on speech production, in addition to subject and item, we also included auditory and visual word frequencies as random intercepts. Consistent with our prediction, we found a significant interaction between our four fixed effects ($\beta = 6.794$, $SE = 2.611$, $t = 2.602$, $p = .01$) specifically for syllables that were rated as a 1 out of 5, indicating proximity to the phonetic boundary.

Interestingly, this interaction occurred specifically for items that were identified as something other than auditory/visual/fusion word initial segment during the syllable block but identified as fusion words during the word block. This interaction indicates, that when integration forms an ambiguous phonetic output, lexical frequency will guide the ultimate phonetic identification. This result converges with our general finding that

fusion word frequency is predictive of fusion word identifications (Experiment 1 & the current experiment). Together, these findings suggest that lexical influences occur at a post-integration stage of processing.

Are There Effects of Auditory and Visual Word Frequency During Integration?

Above we discussed how we continue to find an interaction between auditory and visual word frequency even when controlling for syllable identification. We noted that this could be because of lexical effects on production that are not accounted for by syllable identification, or because of lexical processing during perception.

To address this question we ran an analysis using syllable rating, syllable identification, word identification, and auditory and visual lexical frequency as fixed effects predicting word rating. Random intercepts were set for subject, item, and the lexical frequency of fusion words. This analysis returned two significant interactions indicating that the interaction between auditory and visual word lexical frequency predicted the goodness ratings of non-auditory/visual/fusion words for syllables with non-auditory/visual/fusion syllable identifications that were rated as a 3 out of 5 (β= -3.227, *SE* = 1.573, *t* = -2.051, *p* = .041) and a 4 out of 5 (β= -52.69, *SE* = 23.144, *t* = -2.277, *p* = .023) which is depicted in Figure 3.6.

These two interactions contrast with what we found for fusion word frequency in two important ways. First, where lexical frequency of fusion words interacted with the lowest rated syllables, here we see that the audio-visual lexical frequency interaction was

found for *two* intermediate rated syllables, a pattern more similar to what we would predict for effects during the integration phase of perception.

Second, where the interaction of fusion word frequency coincided with a change of identifications between syllables and words, to form fusion word identifications, here the interaction with lexical frequency did not produce word percepts. This pattern is difficult to attribute to lexical "processing" at any stage of perception; after all lexical processing should correspond with word recovery. Based on this latter observation, we suggest that it is possible that these effects reflect the lexical effects on production noted from Experiment 2.

Discussion

Experiment 3 expands on the findings of Experiments 1 and 2. Recall that a key finding from Experiment 1 was the predictive value of the interaction between auditory and visual word frequencies on word identifications. The results of that experiment were ambiguous concerning whether lexical influences occurred during and/or following multisensory integration. The results of Experiment 2 suggested a third interpretation for Experiment 1. It may be that the influence of lexical information on word identification is attributable, in part, to lexical effects on speech *production.*

The most important finding from Experiment 3 is the four-way interaction between syllable McGurk effect, syllable goodness rating, lexical frequency of visual words, and the lexical frequency of the auditory words. This interaction replicates the finding reported above that McGurk effects are predicted by the interaction between auditory and visual lexical frequencies. However, this analysis localizes that effect to

items that were not identified as either the auditory or visual word, or even fusion words. That lexical processing during perception should promote the identification of words, this finding indicates that the interaction between auditory and visual lexical frequencies is not driven by a perceptual process, and is likely related to lexical effects on speech production.

In contrast, our analysis with fusion word lexical frequency shows that lexical information (word contexts) only changed the goodness ratings of items whose syllables were near the phonetic boundary; a pattern indicative of a post-feature integration/post multisensory integration, phonetic categorization process (Brancazio et., 2003). In conjunction with our results for the auditory and visual word interaction effects noted above, Experiment 3 offers evidence in support of theories that assume lexical information influences categorization, but not integration (e.g. Norris et al., 2003).

General Discussion

In a series of three experiments we investigated how lexical information influences multisensory speech perception. We consider three points during speech perception that have the potential to be influenced by lexical information: during speech production, during multisensory integration, and (or) during phonetic categorization (post integration). It is unlikely that these points during speech perception are discrete and insulated from one another. It is certainly possible that these processes occur in parallel and interact with one another as is assumed by many computational and cascading activation models. However, while there are many such accounts of lexical processing (e.g. McClelland, 2015; McClelland & Elman, 1986; Norris, McQueen, & Cutler, 2000)

and for multisensory integration (e.g. Yuhas, Goldstein, Sejnowski, & Jenkins, 1990; Jantvik, Gustafsson, & Paplinski, 2011; Rahmani, Almasgani, & Syyedsalehi, 2018) few, if any accounts discuss how lexical and multisensory information might interact. In the absence of such a comprehensive account, these three points during speech perception were useful for formulating the questions that motivated this research.

Across all of our experiments there were two consistent findings: 1) auditory speech identifications were best predicted by the *interaction* between auditory and visual lexical frequency and 2) fusion word identifications were best predicted by fusion word lexical frequency. In the following paragraphs we will discuss how these findings inform us about the locus of lexical processing in multisensory speech perception.

The results of Experiment 2 and Experiment 3 both support an account in which lexical information influences multisensory speech perception by affecting speech *production.* Briefly, both of these experiments found the interaction between auditory and visual lexical frequency in the identification of syllables that had been extracted from words. Being only syllables, these stimuli could not support lexical *processing* by the perceiver. Thus this interaction suggests that lexical effects on speech production (e.g. Pluymaekers et al., 2005) are responsible. This conclusion has substantial implications. Methodologically, this finding suggests that future investigations of McGurk effects with words should control for these speech production effects. Theoretically, this finding suggests that accounts for multisensory speech processing should consider the effect of lexical context in both the perceiver who processes the speech and also the talker who produces it.

The results of Experiment 1 and Experiment 3, with McGurk word stimuli both find that fusion word frequency predicts fusion word identifications. As fusion words are not present in the sensory input, this finding indicates that some lexical processing occurs post multisensory integration. That fusion word lexical frequency does not interact with the effects of auditory and visual lexical frequency indicates that the processing of these fusion words does not occur during integration. This conclusion is also consistent with the results of our analysis of the goodness scores using the fixed effect of fusion word frequency. This analysis showed that lexical frequency only affected words with ambiguous syllables, an effect that is predicted by a lexical processing during a post integration phonetic categorization process (i.e. see Allen & Miller, 2001).

Finally, we must consider the evidence for lexical processing during integration. Initially, the interaction between auditory and visual word frequencies predicting auditory and visual word identifications was considered evidence for lexical processing during integration. However, by finding this same audio-visual lexical frequency interaction in the identification of syllables in Experiment 2 (and Experiment 3) suggested that at least part of this effect preceded integration. Finally, in Experiment 3, we analyzed the change in goodness scores between syllables and words. The work of Brancazio et al., (2003) suggested that processing during multisensory integration would produce a broad change in goodness scores.

The interaction between auditory and visual lexical frequency produced significant effects for 2 out of the 5 levels of goodness rating. The prediction of a 'processing during integration' hypothesis would be for interactions across all/most

syllable goodness ratings. As such, while this analysis is more consistent with processing during integration than was the analysis with fusion word frequencies, it still fell quite short of the predicted effect. As such, it should only be interpreted with caution.

Interestingly, this interaction was not associated with auditory and visual word identifications. It is unlikely that lexical processing would produce a bias towards nonwords. Thus it seems that the interaction between auditory and visual lexical frequency, while likely localized to the integration stage, is attributed to the lexical effects on production rather than lexical processing by the perceiver.

We should consider the possibility that the interaction we found in the goodness scores analysis is spurious. It is possible that that interaction reflects something idiosyncratic to goodness scores. It is also possible that we would find different results if we included fusion word lexical frequencies. As fusion word processing occurs post integration, it is conceivable that it would interact with auditory or visual lexical frequency for syllables that were identified as ambiguous with the auditory (or visual) item and the fusion item. However, when we added this term to the analysis, the model failed to converge; likely as a result of too many factors being applied to too few observations.

Ultimately, additional work is needed to understand the goodness scores analysis interactions. However, that the effects with the identification analyses are consistent across all three experiments makes us confident in the conclusions we drew from those analyses. The limitations on the analysis of the goodness scores demands follow up, but do not undermine the conclusions drawn on the basis of the other analyses.

A final limitation worth noting is that this investigation relied on McGurk stimuli. Such stimuli, by design, have incongruent audio-visual structuring. The unnaturalness of these stimuli imposes limitations on the current study, and indeed of most studies of multisensory perception that rely on the McGurk effect (see Alsius et al., 2018 for a review). However, the use of McGurk stimuli allowed the current investigation to compare the lexical processing of auditory and visual stimuli, the benefit of which outweighs these limitations.

Implications For Theories Of Speech Perception

This investigation was, in large part, motivated by the work of Brancazio (2004). Recall that Brancazio (2004) found that there was a bias for participants to identify words from audio word/nonword + visual word/nonword McGurks. He discussed the implications of his results for theories of speech perception including TRACE (McClelland & Elman, 1986) and Merge (Norris, McQueen, & Cutler, 2000) and here we extend that discussion to our own results. Neither Merge or TRACE are models of multisensory perception and for that reason we will also discuss the implications of our results for a prominent theory of multisensory perception, the amodal account. We will also address two other more recent theories that address the interaction of multisensory and lexical processing (Samuel & Lieblich, 2014; Ostrand et al., 2016).

For the amodal account the first stage in the perceptual process is the integration stage. It is assumed that this stage is insulated from all higher level processes, such as lexical processing. For the amodal account, integration is less a cognitive "process" than a result of the congruent structure of the sensory input (Fowler 2004; see also Rosenblum

et al., 2016 for a review). As there is little to no assumed cognitive processing during integration, lexical influences on the McGurk effect should be limited to the categorization process and lexical influences on production. That we find effects of speech production but fail to find evidence for lexical *processing* during integration is consistent with the predictions of this account.

Not being a theory of multisensory speech, Merge does not make specific predictions about integration. However, Merge does offer an explanation for how low level speech information and higher order lexical information can "merge" to inform phoneme identification (Norris et al., 2000). Under this account, both low-level speech information and top-down lexical information feed into a phoneme decision process. An explicit feature of the Merge account is that lexical information can influence phoneme identification though this decision process, but that lexical information *does not* feedback to change the phenomenological experience of the lower level speech units. That is, lexical information will change what category a poor phonetic exemplar is assigned to, but will not make that segment sound like a better exemplar. In this respect, Merge perfectly predicts the results of our analysis of fusion word frequency on word goodness scores; lexical information only affected the categorization.

TRACE is a spreading activation model of speech perception with three levels of processing; featural, phonemic, and lexical (McClelland & Elman, 1986; McClelland et al., 2006). Under this framework (and other cascading activation accounts) activation levels are assumed to be related to lexical factors such as lexical frequency (e.g. McClelland & Rumelhart, 1981). In TRACE, higher level units are activated by the

support from lower level units (e.g. activation of features leads to the activation of associated phonemes), but these higher level units can feedback to bolster the activation of associated lower level units and dampen the activation of competing lower level units. In this way, TRACE holds that lexical processing should feedback to influence the phenomenological experience of segments within word contexts; a contrast to the assumptions of Merge. However, none of our analyses suggest that there was feedback from lexical processes to the integration phase; we find no clear support for TRACE. This may be a result of TRACE not being intended to account for audio-visual speech perception. Other cascading activation computational accounts have been put forward for audio-visual integration (Yuhas et al., 1990; Jantvik et al., 2011; Rahmani et al., 2018) and how these accounts might incorporate effects of lexical factors, such as lexical frequency should be considered in future work.

Recently it has been proposed that multisensory perception and lexical processing are dissociable. Samuel and Lieblich (2014; see also Baart & Samuel, 2015) argue that lexical processing is driven by auditory speech and operates independent of multisensory perception. Similarly, Ostrand et al., (2016) has proposed that lexical processing commences prior to the completion of multisensory integration, and is thus driven, initially, by the unintegrated auditory information.

Thus these two accounts both assume: 1) that lexical processing is preferential to auditory information, and 2) that lexical processing should be independent of multisensory integration. That across all three experiments presented here we only found effects of auditory word frequency through its interaction with visual word frequency,

challenges the primacy of auditory information in speech processing assumed by both of these accounts. Moreover, the fact that this interaction appears to be driven by speech production, not processing from the perceiver is a further challenge to these accounts. Based on the results of the experiments here, namely the reliable effect of fusion word frequency, it seems that lexical processing occurs after integration and *on* the integrated output. This conclusion directly challenges both the Ostrand et al., (2016) and the Samuel and Lieblich (2014) accounts (see also the dissertation general discussion section).

It is worth noting that many of these results can be accounted for by probabilistic models that assume perception is determined by Bayesian causal inference about the environmental causes of sensory inputs. Such accounts have been offered for lexical influences (e.g. Norris & McQueen, 2008; McClelland, 2013) and for multisensory influences on speech perception (e.g. Magnotti & Beauchamp, 2015; 2017; Magnotti, Smith, Salinas, Mays, Zhu, Beauchamp, 2018; see also Shams, 2012 for a non-speech account). Under these accounts, perceptions are determined by combining the probability that a given sensory input corresponds to a particular cause with the probability of that cause given the perceivers prior knowledge (See Shams, 2012). Often the probabilities associated with sensory inputs are assumed to be consistent with confusability of speech segments (e.g. how often a 'ba' is mis-identified as a 'da') in auditory and visual modalities (e.g. Magnotti & Beauchamp, 2017). Under the lexical accounts, prior knowledge is assumed to be quantifiable by factors such as lexical and syllable frequency (e.g. Norris & McQueen, 2008).

These accounts generally hold that higher frequency words have larger priors and thus perception is more likely to be of higher frequency words, which as noted above, is not always the case in our results. However, this might be because our analyses did not control for syllable frequency, a possibility that will need to be explored in future work. It is also worth noting that none of these accounts (to our knowledge) incorporate both interactions between auditory and visual sensory inputs with lexically correlated prior knowledge and these interactions could be key to understanding our present results under a Bayesian framework. Thus the present results, while not readily consistent with current probabilistic accounts, could be integrated with modest extensions of existing models.

In conclusion, in a series of three experiments we investigated when during multisensory speech identification is lexical information processed. We found influences of lexical context on the identification of isolated syllables, which is attributable to lexical effects on the speech production process. Such effects are not only theoretically interesting, but methodologically important for future investigations of lexical effects on perception. We also found evidence for lexical effects on phonetic categorization; an effect that, unlike the effects on production, *can* be attributed to lexical *processing* from the perceiver. However, we were unable to find evidence of lexical processing during multisensory integration; the stage of speech perception responsible for merging cross-sensory inputs does not seem to *process* lexical information even if it is sensitive to the influences of that information on the inputs being merged.

References

Allen, J. S., & Miller, J. L. (2001). Contextual influences on the internal structure of phonetic categories: A distinction between lexical status and speaking rate. *Perception & Psychophysics*, *63*(5), 798–810. http://doi.org/10.3758/BF03194439

Alsius, A., Paré, M., & Munhall, K. G. (2018). Forty years after hearing lips and seeing voices: the McGurk effect revisited. *Multisensory Research*, *31*(1–2), 111–144. http://doi.org/10.1163/22134808-00002565

Baart, M., & Samuel, A. G. (2015). Turning a blind eye to the lexicon: ERPs show no cross-talk between lip-read and lexical context during speech sound processing. *Journal of Memory and Language*, *85*(July). http://doi.org/10.1016/j.jml.2015.06.00

Balota, D. A., & Chumbley, J. I. (1985). The locus of word-frequency effects in the pronunciation task: Lexical access and/or production? *Journal of Memory and Language*, *24*, 89–106.

Barutchu, A., Crewther, S. G., Kiely, P., Murphy, M. J., & Crewther, D. P. (2008). When /b/ill with /g/ill becomes /d/ill: Evidence for a lexical effect in audiovisual speech perception. *European Journal of Cognitive Psychology*, *20*(1), 1–11. http://doi.org/10.1080/09541440601125623

Bernstein, L. E., Eberhardt, S. P., & Auer, E. T. (2014). Audiovisual spoken word training can promote or impede auditory-only perceptual learning: Prelingually deafened adults with late-acquired cochlear implants versus normal hearing adults. *Frontiers in Psychology*, *5*, 1–20. http://doi.org/10.3389/fpsyg.2014.00934

Bertelson, P., Vroomen, J., & De Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science*, *14*(6), 592–597. http://doi.org/10.1046/j.0956-7976.2003.psci_1470.x

Brancazio, L. (2004). Lexical influences in audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(3), 445–463. http://doi.org/10.1037/0096-1523.30.3.445

Brancazio, L., & Miller, J. L. (2005). Use of visual information in speech perception: Evidence for a visual rate effect both with and without a McGurk effect. *Perception and Psychophysics*. http://doi.org/10.3758/BF03193531

Brancazio, L., Miller, J. L., & Paré, M. A. (2003). Visual influences on the internal structure of phonetic categories. *Perception & Psychophysics*, *65*(4), 591–601. http://doi.org/10.3758/BF03194585

Colombo, L., Pasini, M., & Balota, D. A. (2006). Dissociating the influence of familiarity and meaningfulness from word frequency in naming and lexical decision performance, (6), 1312–1324.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General*, *134*(2), 222–241. http://doi.org/2005-04168-006 [pii]\n10.1037/0096-3445.134.2.222

Drouin, J. R., Theodore, R. M., & Myers, E. B. (2016). Lexically guided perceptual tuning of internal phonetic category structure. *The Journal of the Acoustical Society of America*, *140*(4), EL307-EL313. http://doi.org/10.1121/1.4964468

Evans, B. G., & Iverson, P. (2004). Vowel normalization for accent: An investigation of best exemplar locations in northern and southern British English sentences. *The Journal of the Acoustical Society of America*, *115*(1), 352–361. http://doi.org/10.1121/1.1635413

Forster, K. I., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 680–698. http://doi.org/10.1037/0278-7393.10.4.680

Fowler, C. A. (2004). Speech as a supramodal or amodal phenomenon. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *Handbook of Multisensory Processes* (pp. 189–201). Cambridge.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology. Human Perception and Performance*, *6*(1), 110–125. http://doi.org/10.1037/0096-1523.6.1.110

Grant, K. W., & Seitz, P. F. P. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, *108*(3), 1197–1208. http://doi.org/10.1121/1.422512

Green, K. P., & Miller, J. L. (1985). On the role of visual rate information in phonetic perception. *Perception & Psychophysics*, *38*(3), 269–276. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/4088819

Griffin, Z. M., & Bock, J. K. (1998). Constraint, word frequency, and levels of processing in spoken word production. *Journal of Memory and Language*, *38*(38), 313–338.

Hirsh, I. J., Reynolds, E. G., & Joseph, M. (1954). Intelligibility of different speech materials. *The Journal of the Acoustical Society of America*, *26*(4), 530–538. http://doi.org/10.1121/1.1907370

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *Journal of the Acoustical Society of America*, *97*(1), 553–562. http://doi.org/10.1121/1.412280

Jantvik, T., Gustafsson, L., & Papliński, A. P. (2011). A self-organized artificial neural network architecture for sensory integration with applications to letter-phoneme integration. *Neural Computation*, *23*(8), 2101–2139.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2000). Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure*. Amsterdam.

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics*, *24*(3), 253–7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/704285

Magnotti, J. F., & Beauchamp, M. S. (2015). The noisy encoding of disparity model of the McGurk effect. *Psychonomic Bulletin & Review*, *22*(3), 701–709. http://doi.org/10.3758/s13423-014-0722-2

Magnotti, J. F., & Beauchamp, M. S. (2017). A Causal Inference Model Explains Perception of the McGurk Effect and Other Incongruent Audiovisual Speech. *PLoS Computational Biology*, *13*(2), 1–15. http://doi.org/10.1371/journal.pcbi.1005229

Magnotti, J. F., Smith, K. B., Salinas, M., Mays, J., Zhu, L. L., & Beauchamp, M. S. (2018). A causal inference explanation for enhancement of multisensory integration by co-articulation. *Scientific Reports*, *8*(1), 18032. http://doi.org/10.1038/s41598-018-36772-8

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1–2), 71–102. http://doi.org/10.1016/0010-0277(87)90005-9

McClelland, J. L. (2013). Integrating probabilistic models of perception and interactive neural networks: A historical and tutorial review. *Frontiers in Psychology*, *4*(AUG), 1–25. http://doi.org/10.3389/fpsyg.2013.00503

McClelland, J. L. (2015). Capturing gradience, continuous change, and quasi-regularity in sound, word, phrase, and meaning. *The Handbook of Language Emergence*, 53–80.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. http://doi.org/10.1016/0010-0285(86)90015-0

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, *88*(5), 375–407. http://doi.org/10.1016/S0022-0728(02)01421-3

McClelland, J. L., Mirman, D., & Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Sciences*, *10*(8), 363–369. http://doi.org/10.1016/j.tics.2006.06.007

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.

Miller, G. A, Heise, G. A, & Lighten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *The Journal of Experimental Psychology*, *41*(5), 329–335.

Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics*, *46*(6), 505–512. http://doi.org/10.3758/BF03208147

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395. http://doi.org/10.1037/0033-295X.115.2.357

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*, 204–238.

Norris, D., McQueen, J., & Cutler, A. (2000). Merging information in speech processing: Feedback is never necessary. *Behavioral and Brain Sciences*, *23*(3), 299–370.

Ostrand, R., Blumstein, S. E., Ferreira, V. S., & Morgan, J. L. (2016). What you see isn't always what you get: Auditory word signals trump consciously perceived words in lexical access. *Cognition*, *151*, 96–107. http://doi.org/10.1016/j.cognition.2016.02.019

Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2005). Lexical frequency and acoustic reduction in spoken Dutch. *The Journal of the Acoustical Society of America*, *118*(4), 2561–2569. http://doi.org/10.1121/1.2011150

Pollack, I., Rubenstein, H., & Decker, L. (1960). Analysis of incorrect responses to an unknown message set. *The Journal of the Acoustical Society of America*, *32*(4), 454–457. http://doi.org/10.1121/1.1908097

Rahmani, M. H., Almasganj, F., & Seyyedsalehi, S. A. (2018). Audio-visual feature fusion via deep neural networks for automatic speech recognition. *Digital Signal Processing*, *82*, 54–63. http://doi.org/10.1016/j.dsp.2018.06.004

Rosenblum, L. D., Dorsi, J., & Dias, J. W. (2016). The impact and status of Carol Fowler's Supramodal Theory of Multisensory Speech Perception. *Ecological Psychology*, *28*(4), 262–294. http://doi.org/10.1080/10407413.2016.1230373

Rosenblum, L. D., & Saldaña, H. M. (1992). Discrimination tests of visually influenced syllables. *Perception & Psychophysics*, *52*(4), 461–473. http://doi.org/10.3758/BF03206706

Sams, M., Manninen, P., Surakka, V., Helin, P., & Kättö, R. (1998). McGurk effect in Finnish syllables, isolated words, and words in sentences: Effects of word meaning and sentence context. *Speech Communication*, *26*(1–2), 75–87. http://doi.org/10.1016/S0167-6393(98)00051-X

Samuel, A. G., & Lieblich, J. (2014). Visual speech acts differently than lexical context in supporting speech perception. *Journal of Experimental Psychology. Human Perception and Performance*, *40*(4), 1479–90. http://doi.org/10.1037/a0036656

Samuel, A., & Kat, D. (1996). Early levels of analysis of speech. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(3), 676–694.

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(1), 1–17. http://doi.org/10.1037/0096-1523.3.1.1

Schilling, H. E. H., Rayner, K., & Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory and Cognition*, *26*(6), 1270–1281. http://doi.org/10.3758/BF03201199

Shams, L. (2012). Early integration and bayesian causal inference in multisensory perception. In M. M. Murray & M. T. Wallace (Eds.), *The neural basis of multisensory processes* (pp. 217–229). Boca Raton, FL: CRC Press/Taylor & Francis. http://doi.org/10.1152/jn.00497.2006

Strand, J., Cooperman, A., Rowe, J., & Simenstad, A. (2014). Individual differences in susceptibility to the McGurk effect: Links with lipreading and detecting audiovisual incongruity. *Journal of Speech, Language and Hearing Reseach*, *57*, 2322–2331. http://doi.org/10.1044/2014

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*(2), 212–215.

Ujiie, Y., Asai, T., & Wakabayashi, A. (2018). Individual differences and the effect of face configuration information in the McGurk effect. *Experimental Brain Research*, *0*(0), 0. http://doi.org/10.1007/s00221-018-5188-4

Yuhas, B. P., Goldstein, M. H., Sejnowski, T. J., & Jenkins, R. E. (1990). Neural network models of sensory integration for improved vowel recognition. *Proceedings of the IEEE*, *78*(10), 1658–1668.

Table 3.1

Stimulus		McGurk Identification			
Audio	Video	Audio	Visual	Top Fusion	Other Fusions
Bury	Very	0.20	0.40	'F' (0.4)	
Banish	Vanish	0.12	0.82		
Ban	Van	0.07	0.67	'F' (0.27)	
Ballad	Valid	0.17	0.83		
Ballet	Valet	0.27	0.73		
Bail	Vial	0.00	0.50	'F' (0.47)	
Bane	Vain	0.07	0.78	'F' (0.09)	
Pug	Tug	0.22	0.39	'H' (0.36)	'Th' (0.03)
Pot	Tot	0.05	0.72	'C' (0.15)	'Th' (0.05) 'H' (0.03)
Pole	Toll	0.15	0.78	'C' (0.03)	'W' (0.02)
Pie	Tie	0.17	0.68	'Th' (0.15)	
Pest	Test	0.08	0.90		
Part	Tart	0.22	0.73	'Th' (0.05)	
Pad	Tad	0.12	0.80	'Th' (0.05)	'H' (0.02) 'A' (0.02)
Mode	Node	0.27	0.62	'L' (0.12)	
Mob	Knob	0.52	0.47	'O' (0.02)	
Mix	Nix	0.23	0.77		
Mine	Nine	0.67	0.28	'L' (0.05)	
Mill	Nil	0.43	0.57		
Might	Night	0.60	0.38		
Mice	Nice	0.65	0.30	'L' (0.05)	
Map	Nap	0.03	0.97		
Maim	Name	0.08	0.90	'A' (0.02)	
Mail	Nail	0.17	0.83		
Me	KNee	0.13	0.87		
Primp	Crimp	0.87	0.13		
Pod	Cod	0.25	0.05		
Buy	Guy	0.63	0.00	'Th' (0.3)	'D' (0.03) 'i' (0.02)
Butt	Gut	0.78	0.00		
Bun	Gun	0.80	0.00	'D' (0.05)	'F' (0.02)
Bum	Gum	0.62	0.00	'Th' (0.38)	
Bore	Gore	0.18	0.07	'Th' (0.52)	'd' (0.18) 'f' (0.05)
Bill	Dill	0.42	0.25		'T' (0.03) 'F' (0.02)
Bid	Did	0.43	0.35		'H' (0.02)
Bet	Debt	0.37	0.28	'V' (0.07)	
Bell	Dell	0.52	0.18		'T' (0.05) 'F' (0.02)
Beer	Dear	0.58	0.20	'F' (0.03)	'T' (0.02) 'F' (0.03)

Bean	Dean	0.42	0.25		
Bead	Deed	0.28	0.22	'F' (0.13)	
Bay	Day	0.35	0.03	'Th' (0.6)	'H' (0.02)
Bank	Dank	0.27	0.02	'Th' (0.7)	
Bait	Date	0.38	0.23	'F' (0.18)	
Bad	Dad	0.47	0.08		'F' (0.03)
Bowel	Vowel	0.07	0.75	'F' (0.17)	'T' (0.02)
Bow	Vow	0.12	0.78	'F' (0.1)	
Boat	Vote	0.15	0.83	'T' (0.02)	
Bolt	Volt	0.10	0.82		
Bowl	Vole	0.33	0.50	'F' (0.17)	
Bile	Vial	0.00	0.83	'F' (0.15)	
Bigger	Vigor	0.35	0.42	'F' (0.23)	
Buy	Vie	0.42	0.43	'P' (0.07)	'Th' (0.02) 'H' (0.02)
Bet	Vet	0.08	0.85		
Best	Vest	0.07	0.82	'F' (0.12)	
Burst	Versed	0.15	0.05	'F' (0.8)	
Bent	Vent	0.08	0.88	'F' (0.03)	
Bender	Vendor	0.10	0.78	'F' (0.12)	
Bending	Vending	0.22	0.75	'F' (0.03)	
Beer	Veer	0.08	0.17	'F' (0.73)	
Bat	Vat	0.07	0.40	'F' (0.37)	'Th' (0.17)
Base	Vase	0.07	0.60	'F' (0.33)	

Table 3.1 shows identification data for Experiment 1.

Table 3.2

Audio Perception = LF(Audio)*LF(Visual)+[(Subject)+(Item)]

	β	SE	df	t	p
Intercept	0.600	0.262	61.753	2.292	0.0254
LF(A)	-0.176	0.088	59.561	-2.005	0.0495
LF(V)	-0.129	0.090	59.561	-1.419	0.161
LF(A)xLF(V)	0.0671	0.029	59.561	2.284	0.026

Table 3.2 displays the results of the analysis of the effect of auditory and visual word frequency on auditory word identifications for Experiment 1.

Table 3.3

Audio Perception = LF(Audio)*LF(Visual)*LF(Fusion)+[(Subject)+(Item)]

	β	SE	df	t	p
(Intercept)	1.102	1.08	38.661	1.021	0.314
LF(A)	-0.373	0.337	38.57	-1.106	0.276
LF(V)	-0.293	0.374	38.57	-0.782	0.439
LF(F)	-0.366	0.371	38.57	-0.985	0.331
LF(A) x LF(V)	0.121	0.113	38.57	1.063	0.294
LF(A) x LF(F)	0.126	0.116	38.57	1.084	0.285
LF(V) x LF(F)	0.108	0.129	38.57	0.836	0.408
LF(A) x LF(V) x LF(F)	-0.033	0.04	38.57	-0.837	0.408

Visual Perception = LF(Audio)*LF(Visual)*LF(Fusion)+[(Subject)+(Item)]

	β	SE	df	t	p
(Intercept)	-1.963	1.474	38.963	-1.332	0.1906
LF(A)	0.966	0.461	38.941	2.098	0.0425
LF(V)	1.009	0.511	38.941	1.974	0.0555
LF(F)	0.787	0.507	38.941	1.553	0.1285
LF(A) x LF(V)	-0.35	0.155	38.941	-2.259	0.0296
LF(A) x LF(F)	-0.301	0.159	38.941	-1.901	0.0647
LF(V) x LF(F)	-0.308	0.176	38.941	-1.745	0.0889
LF(A) x LF(V) x LF(F)	0.101	0.054	38.941	1.853	0.0715

Table 3.3 displays the results of the analyses of the effect of auditory and visual and fusion word frequency on auditory and visual word identifications for Experiment 1.

Table 3.4

	McGurk: Proportion			Cong: Proportion		McGurk: Rating			Cong: Rating	
Family	Audio	Visual	Fusion	CongA	CongV	Audio	Visual	Fusion	CongA	CongV
Cod	0.11	0.00	0.00	0.91	0.64	2		1.875	2.4	4.286
Crimp	0.89	0.00	0.00	0.91	0.73	4.125		1	3	2
Dad	0.11	0.33	0.00	0.91	0.91	2	2.333	3	3.4	3
Dank	0.22	0.44	0.22	1.00	1.00	3.5	2	2.667	2.3634	1.909
Date	0.00	0.44	0.11	0.91	0.82		2.5	2.4	2.8	2.556
Day	0.00	0.67	0.00	0.91	1.00		2.667	3.333	3.2	2.818
Dean	0.00	0.89	0.00	1.00	1.00		4.25	4	4	4.091
Dear	0.22	0.56	0.11	0.91	1.00	2.5	3.6	3	3	3.364
Debt	0.00	0.56	0.00	0.82	1.00		2.2	3	2.556	4.364
Deed	0.00	0.78	0.11	1.00	0.91		3.429	3.5	3.818	4
Dell	0.11	0.44	0.44	0.91	1.00	2	2.25	2.5	2.8	2.455
Did	0.00	0.67	0.00	1.00	0.82		3.167	3. 333	3.545	4. 333
Dill	0.11	0.78	0.00	0.91	1.00	2	2.429	4	2.9	2.364
Gore	0.22	0.11	0.33	0.82	0.91	2.5	3	2.667	2.778	2.2
Gum	0.11	0.00	0.22	0.91	0.82	3		2.375	3.3	2.778
Gun	0.33	0.00	0.33	1.00	0.91	3.333		2.333	2.818	3.1
Gut	0.56	0.00	0.00	0.91	1.00	2		2.5	2	3.091
Guy	0.44	0.00	0.00	1.00	0.91	4		3.4	3.364	3.2
Knee	0.11	0.89	0.00	0.91	0.18	4	4		2.5	4.5
Knob	0.56	0.33	0.00	1.00	1.00	3.6	1.667	1	3	2.273
Nail	0.00	1.00	0.00	1.00	0.82		4		2.818	2.556
Name	0.00	1.00	0.00	1.00	1.00		4.111		4	3.818
Nap	0.00	1.00	0.00	0.91	0.91		2.667		1.9	2.7
Nice	0.22	0.67	0.00	0.82	1.00	3	3.333	1	2.889	2.364

Night	0.33	0.56	0.00	0.91	1.00	3	2.6	1	2.4	3.636
Nil	0.00	1.00	0.00	0.91	0.91		3.778		3.9	1.9
Nine	0.56	0.44	0.00	1.00	1.00	3	3		3	3.364
Nix	0.00	1.00	0.00	0.91	0.91		4.556		3.5	2.4
Node	0.22	0.67	0.00	0.91	1.00	4	3.833	1	3.2	3.909
Tad	0.00	1.00	0.00	1.00	1.00		3.778		4.273	3.636
Tart	0.33	0.67	0.00	0.91	0.91	2	2.833		2.9	2.5
Test	0.11	0.89	0.00	1.00	1.00	3	3		2.091	3.182
Tie	0.11	0.89	0.00	1.00	0.91	3	3.125		3.364	4.1
Toll	0.11	0.56	0.11	0.91	1.00	2	1.8	3.667	3.2	3.818
Tot	0.22	0.78	0.00	1.00	1.00	4	4		4.091	3.818
Tug	0.44	0.56	0.00	0.91	1.00	2.5	3.4		3.3	4
Vain	0.11	0.33	0.22	0.82	0.91	1	2.667	3	2.556	3.9
Valet	0.11	0.44	0.00	0.91	0.82	3	3.25	3	2.4	4.444
Valid	0.33	0.44	0.00	1.00	0.82	3.333	4	3	2.727	3.333
Van	0.22	0.33	0.33	0.91	0.82	2	4	3.5	2.6	3.667
Vanish	0.22	0.56	0.00	0.45	0.82	3	2.4	4.5	1.4	3.333
Vase	0.11	0.33	0.44	0.91	0.82	2	4.333	4.2	3.5	3.556
Vat	0.11	0.33	0.33	0.82	0.55	2	3.667	3.2	2.556	2.5
veil	0.22	0.11	0.56	0.82	0.73	3	5	4.167	3.222	4.125
Vending	0.22	0.44	0.11	0.91	0.82	3.5	2.75	3	3.1	2.556
Vendor	0.33	0.56	0.00	0.91	0.45	3	2.8	4	2.4	1.8
Vent	0.11	0.44	0.22	1.00	0.82	2	2.75	2.25	2.182	3.111
Versed	0.22	0.56	0.22	1.00	0.73	2.5	3.6	4.5	2.636	3.625
Very	0.33	0.44	0.00	0.91	0.45	4.333	4.25	4.5	3	1.6
Vest	0.22	0.22	0.22	0.73	1.00	2.5	3	2.8	2.75	3.727
Vial	0.22	0.33	0.33	1.00	1.00	4	3.667	4	3.182	3.273
Vigor	0.44	0.22	0.22	0.91	0.55	3.25	4	2.667	2.4	1.833

Vole	0.11	0.56	0.33	1.00	0.82	4	4	3.333	3.455	3.556
Volt	0.22	0.67	0.00	0.91	0.91	2	2.667	3	2.9	1.9
Vote	0.11	0.56	0.11	0.82	0.91	3	2.8	2.333	3.333	3.2
Vow	0.22	0.44	0.22	0.91	1.00	1	3.5	2.667	2.6	3.182
Vowel	0.33	0.33	0.11	0.82	0.91	3.333	3.667	3	2.333	2.9

Table 3.4 shows identification and goodness ratings for Experiment 2.

Table 3.5

Audio Perception = LF(Audio)*LF(Visual)*LF(Fusion)+[(Subject)+(Item)]

	β	SE	z	p
(Intercept)	-2.046	6.442	-0.318	0.751
LF(A)	-0.013	1.986	-0.007	0.995
LF(V)	0.867	2.369	0.366	0.715
LF(F)	1.070	2.256	0.474	0.635
LF(A) x LF(V)	-0.280	0.715	-0.392	0.695
LF(A) x LF(F)	-0.365	0.704	-0.519	0.604
LF(V) x LF(F)	-0.682	0.817	-0.835	0.404
LF(A) x LF(V) x LF(F)	0.237	0.252	0.939	0.348

Audio Perception = LF(Fusion)+[(Subject)+(Item)]

	Beta	SE	z	p
(Intercept)	-2.041	0.523	-3.9	<.01
LF(F)	0.058	0.153	0.381	0.703
LF(A) x LF(V)	0.560	0.089	6.327	<.01

Audio Perception = LF(Audio)*LF(Visual)+[(Subject)+(Item)]

	Beta	SE	z	p
(Intercept)	3.157	0.796	3.964	<.05
LF(A)	-1.791	0.271	-6.606	<.05
LF(V)	-1.542	0.274	-5.618	<.05
LF(A) x LF(V)	0.560	0.089	6.327	<.05

Table 3.5 displays the results of the multiple analyses of the effect of auditory and visual and fusion word frequency on auditory syllable identifications for Experiment 2.

Table 3.6a

Goodness Rating = LF(Audio)*LF(Visual)+[(Subject)+(Item)]

	β	SE	df	t	p
(Intercept)	4.021	0.674	60.668	5.966	<.05
LF(A)	-0.358	0.222	55.882	-1.614	0.112
LF(V)	-0.139	0.230	55.483	-0.631	0.531
LF(A) x LF(V)	0.067	0.072	54.792	0.94	0.351

Table 3.6a displays the results of the analysis of the effect of auditory and visual and fusion word frequency on auditory syllable goodness scores for Experiment 2.

Table 3.6b

Goodness Rating =LF(A)*LF(V)*Congruent Auditory rating *Congruent Visual rating+[(Subject)+(Item)]

	Beta	SE	df	t	p
(Intercept)	-4.13E+01	2.00E+01	5.63E+01	-2.07	0.04307
CongA	1.58E+01	6.64E+00	5.63E+01	2.378	0.02083
CongV	1.42E+01	5.76E+00	5.63E+01	2.467	0.01668
LF(A)	1.40E+01	6.51E+00	5.63E+01	2.145	0.03629
LF(V)	2.01E+01	9.68E+00	5.63E+01	2.08	0.04206
CongA:CongV	-4.88E+00	1.87E+00	5.63E+01	-2.606	0.01169
CongA:LF(A)	-5.24E+00	2.21E+00	5.63E+01	-2.372	0.02112
CongV:LF(A)	-4.55E+00	1.84E+00	5.63E+01	-2.476	0.01633
CongA:LF(V)	-6.80E+00	3.23E+00	5.63E+01	-2.106	0.0397
CongV:LF(V)	-6.23E+00	2.58E+00	5.63E+01	-2.417	0.01892
LF(A):LF(V)	-6.63E+00	3.32E+00	5.63E+01	-1.995	0.0509
CongA:CongV:LF(A)	1.66E+00	6.13E-01	5.63E+01	2.709	0.00891
CongA:CongV:LF(V)	2.06E+00	8.43E-01	5.63E+01	2.444	0.01769
CongA:LF(A):LF(V)	2.34E+00	1.13E+00	5.63E+01	2.079	0.04223
CongV:LF(A):LF(V)	2.01E+00	8.49E-01	5.63E+01	2.37	0.02123
CongA:CongV:LF(A):LF(V)	-7.02E-01	2.86E-01	5.63E+01	-2.457	0.01712

Table 3.6b displays the results of the analysis of the effect of auditory and visual word frequency and congruent auditory and visual goodness ratings on auditory syllable goodness scores for Experiment 2.

Table 3.6c

Goodness Rating =LF(A)* Congruent Auditory rating *Congruent Visual rating +[(Subject)+(Item)]

	Beta	SE	df	t	p
(Intercept)	6.36312	6.0587	56.36056	1.05	0.298
CongA	-0.41957	2.04455	56.31338	-0.205	0.838
CongV	-1.02968	1.78221	56.3134	-0.578	0.566
LF(A)	-1.78445	2.03298	56.31342	-0.878	0.384
CongA:CongV	0.20061	0.59036	56.31339	0.34	0.735
CongA:LF(A)	0.35414	0.69514	56.31343	0.509	0.612
CongV:LF(A)	0.43229	0.59133	56.31344	0.731	0.468
CongA:CongV:LF(A)	-0.08664	0.1992	56.31344	-0.435	0.665

Table 3.6d

Goodness Rating =LF(V)*Congruent Auditory rating *Congruent Visual ratin+[(Subject)+(Item)]

	Beta	SE	df	t	p
(Intercept)	-1.02625	3.97334	56.45406	-0.258	0.797
CongA	1.0676	1.32111	56.34428	0.808	0.422
CongV	0.86821	1.12704	56.34431	0.77	0.444
LF(V)	1.16604	2.08143	56.34428	0.56	0.578
CongA:CongV	-0.15623	0.36802	56.34432	-0.425	0.673
CongA:LF(V)	-0.27941	0.69393	56.34429	-0.403	0.689
CongV:LF(V)	-0.32784	0.55215	56.34427	-0.594	0.555
CongA:CongV:LF(V)	0.06787	0.18169	56.34428	0.374	0.71

Tables 3.6c and d displays the results of the analysis of the effect of auditory and visual word frequency and congruent auditory and visual goodness ratings on auditory syllable goodness scores for Experiment 2.

Table 3.7

	Syllable: Proportion			Word: Proportion			Syllable: Rating				Word: Rating			
Family	Audio	Visual	Fusion	Audio	Visual	Fusion	Audio	Visual	Fusion	nonLex	Audio	Visual	Fusion	nonLex
Cod	0.2	0	0	0.25	0.05	0	1.75			3	3.6	4		4.214
Crimp	0.95	0	0	0.9	0.1	0	3.211			1	4.556	3		
Dad	0	0.4	0.1	0.25	0.4	0		3.25	3.5	3	3.2	3.875		3
Dank	0.1	0.5	0.1	0.1	0.3	0.35	2.5	2	2.5	2.167	4	3.5	4.571	4.2
Date	0.2	0.45	0.1	0	0.6	0.15	3.25	2.889	2.5	2		3.833	3.333	4
Day	0.15	0.4	0.1	0.05	0.3	0.35	4	2.875	3.5	1.714	5	4	4.429	4.167
Dean	0.05	0.55	0	0	0.7	0	3	3.636		3.125		3.786		4
Dear	0.1	0.6	0	0.2	0.55	0.05	1.5	3.333		2.667	4	3.727	4	4
Debt	0	0.55	0	0.1	0.6	0.05		3.273		2.889	3.5	4.333	3	3.4
Deed	0.1	0.45	0.05	0.1	0.5	0.1	4	3.778	2	3.5	3	3.8	4.5	4
Dell	0.05	0.55	0.35	0.35	0.35	0.15	3	2.818	2.286	1	3.857	3.143	2	3.667
Did	0	0.65	0	0.15	0.7	0		3.75		3.143	4	3.857		3.667
Dill	0.1	0.5	0	0.1	0.55	0	2.5	2.5		2.125	2	4		4
Gore	0.15	0.05	0.05	0.15	0	0.25	3.33	2	3	3	4		5	4.083
Gum	0.15	0	0.2	0.15	0	0.3	4		1.75	2.538	4		4.333	4.273
Gun	0.35	0	0.3	0.5	0	0.3	3.143		3.5	2.571	3.7		3.167	3.75
Gut	0.25	0	0	0.6	0	0	2			2.2	4.083			3.625
Guy	0.6	0	0.05	0.45	0	0.2	3.167		4	3.571	4.778		4	3.286
Knee	0.15	0.5	0	0.05	0.6	0	4	3.7		2.714	5	4.75		4.714
Knob	0.55	0.4	0	0.65	0.35	0	2.727	2.625		1	3.923	3.857		
Nail	0.1	0.85	0	0.1	0.9	0	3.5	4.058		2	4.5	4.778		
Name	0.1	0.8	0	0	1	0	3	4.438		2		4.6		
Nap	0	0.95	0	0	1	0		2.947		2		4.7		
Nice	0.4	0.55	0	0.45	0.55	0	3	3.455		2	4.333	4.364		
Night	0.5	0.45	0	0.6	0.35	0.05	2.7	2.444		1	3.833	4.714		

Nil	0.15	0.8	0	0.05	0.95	0	3.667	4.313		1	2	4.316			
Nine	0.6	0.35	0	0.75	0.25	0	3.833	3.1429		1	3.8	4.4			
Nix	0.05	0.9	0	0.05	0.95	0	3	3.944		1	5	4.684			
Node	0.25	0.7	0	0.45	0.5	0.05	2.4	4.571		2	3.667	3.8	4		
Tad	0.05	0.95	0	0.05	0.95	0	2	3.947			4	4.263			
Tart	0.15	0.85	0	0.35	0.65	0	2	1.824			4.143	4.385			
Test	0.2	0.8	0	0.25	0.75	0	1.5	3.188			4.4	4.667			
Tie	0.3	0.6	0	0.3	0.7	0	2.167	4.083		4.5	4.167	4.357			
Toll	0.25	0.7	0	0.25	0.75	0	2.2	3.071		5	3	3.467			
Tot	0.1	0.85	0	0.25	0.7	0	3.5	4.118		1	4	4.071		2	
Tug	0.5	0.35	0	0.65	0.3	0	3.4	3		3	4.538	4.667		5	
Vain	0.05	0.35	0.35	0.05	0.45	0.2	1	2.571	2.143	3.2	3	4.444	4.75	3.167	
Valet	0.1	0.65	0	0.35	0.65	0	3.5	3.077		3.6	4.714	4.385			
Valid	0.05	0.5	0	0.1	0.75	0.05	1	2.7		2	4.5	4.533		4	
Van	0.1	0.35	0.35	0	0.7	0.1	3	2.571	3.286	3.75		4.429	5	4.75	
Vanish	0.2	0.45	0	0.1	0.8	0	2.5	2.222		2.286	4	4.5		5	
Vase	0.05	0.5	0.3	0	0.6	0.4	4	2.9	4.5	3.333		4.333	4.5		
Vat	0.15	0.5	0.3	0.1	0.4	0.4	3.333	2.7	4	1	3	3.75	5	3	
veil	0.05	0.3	0.35	0.05	0.35	0.5	5	2.833	4.143	2.5	4	3.714	4.7	4	
Vending	0.2	0.4	0.3	0.15	0.7	0.1	3.5	3.25	3.5	3.5	4	4.429	4	1	
Vendor	0.2	0.45	0.3	0.1	0.55	0.25	3.25	2.222	2.667	2	4.5	4.636	4.2	5	
Vent	0.1	0.3	0.3	0.05	0.95	0	4	1.833	1.667	3	5	4.737			
Versed	0.15	0.4	0.35	0.15	0.1	0.75	1	2.75	3	3	5	5	4.133		
Very	0.15	0.6	0.2	0.15	0.55	0.25	3	3.25	3.75	3	4.333	4	4.6	5	
Vest	0.25	0.25	0.35	0.05	0.95	0	1.8	3	3.714	2	5	4.368			
Vial	0.15	0.4	0.35	0.15	0.45	0.3	4.667	2.75	3	3	4.333	4.889	3.833	5	
Vigor	0.25	0.45	0.3	0.35	0.35	0.3	2.4	2.444	2.5		4.286	3.857	4.5		
Vole	0.15	0.35	0.45	0.15	0.5	0.35	3.667	3.429	3	4	2.333	4	4.571		

Volt	0.1	0.55	0	0.05	0.85	0	2.5	3		2.286		1	3.706		4.5
Vote	0.15	0.45	0.1	0.2	0.65	0.05	3	2.444	2	2.667		4.5	4.462	5	3.5
Vow	0.2	0.35	0.35	0.1	0.85	0.05	1.5	2.429	3.857	2		4	4.1765	4	
Vowel	0.2	0.45	0.3	0.1	0.75	0.05	3.75	2.111	2.667	2		5	4.667	5	5

Table 3.7 shows identification and goodness ratings for Experiment 3.

Table 3.8

Syllable McGurks = LF(Audio)*LF(Visual)+[(Subject)+(Item)]

	β	SE	z	p
(Intercept)	-3.379	1.6279	-2.076	0.03792
LF(A)	1.9376	0.5573	3.477	0.000507
LF(V)	1.714	0.5571	3.077	0.002093
LF(A) x LF(V)	-0.61	0.1828	-3.338	0.000844

Table 3.8 displays the results of the analysis of the effect of auditory and visual word frequency on auditory syllable identifications for Experiment 3.

Figure 3.1

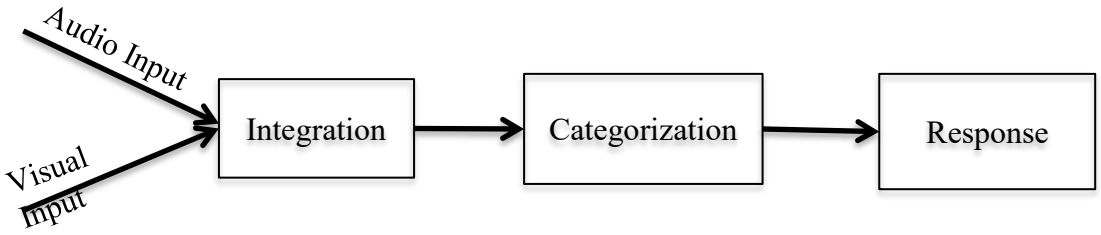


Figure 3.1 depicts the basic framework for understanding audio-visual speech identification; adapted from discussion provided by Brancazio (2004).

Figure 3.2

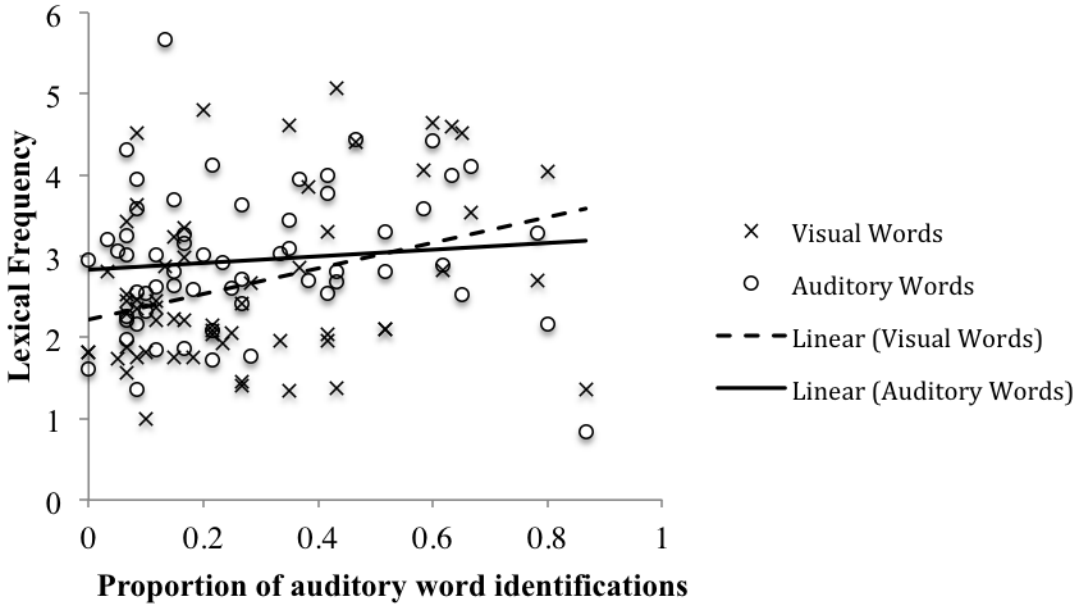


Figure 3.2 shows the relationship between auditory and visual word identification by auditory and visual word lexical frequency in Experiment 1.

Figure 3.3

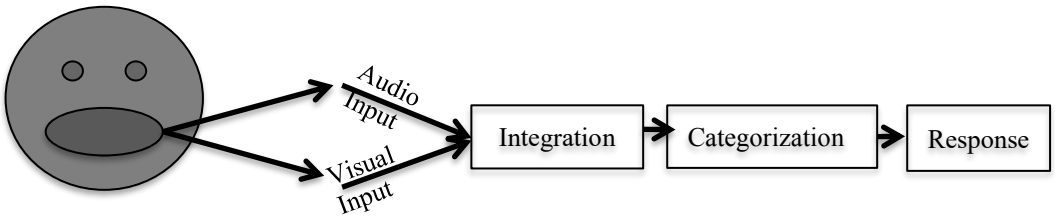


Figure 3.3 depicts our addition to the Bracazio framework, adding a production stage indicated here by the illustrated face.

Figure 3.4

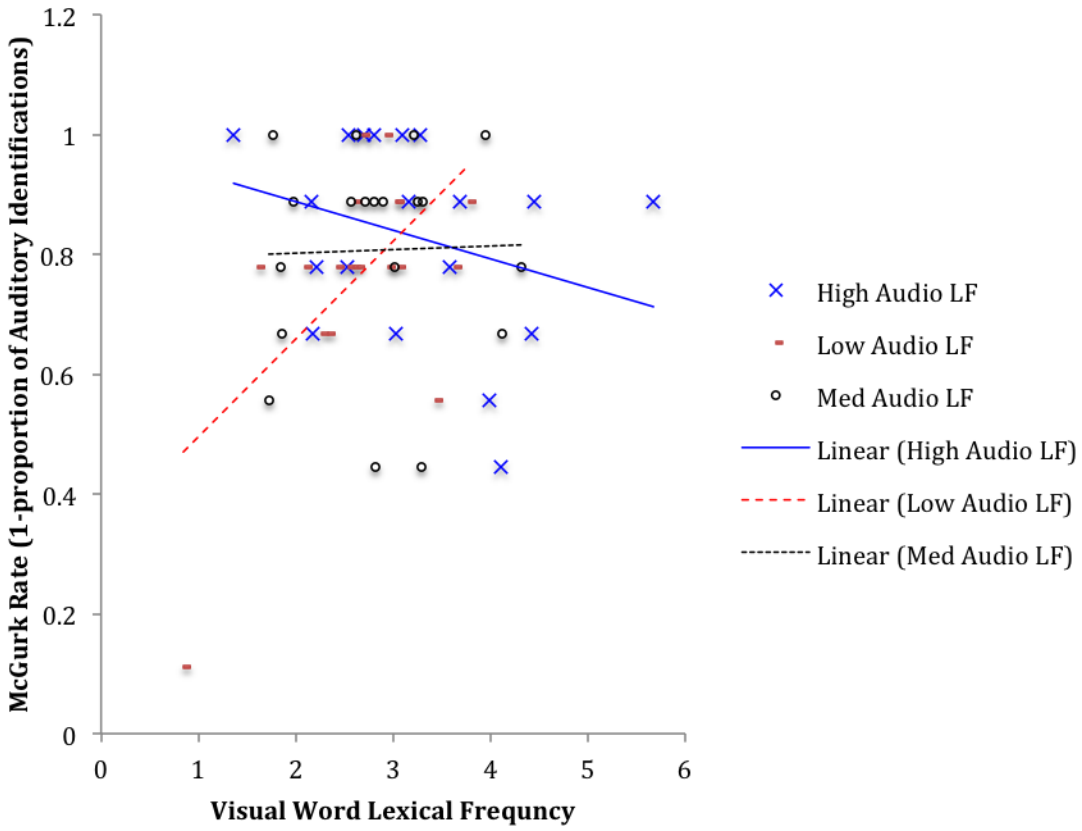
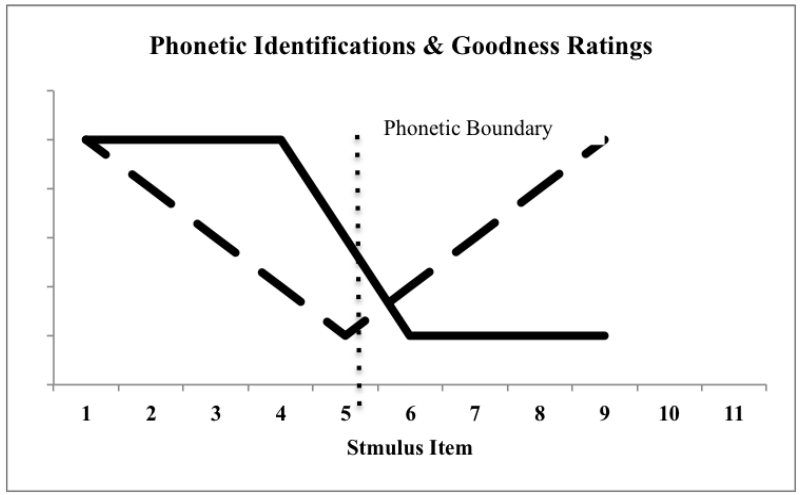
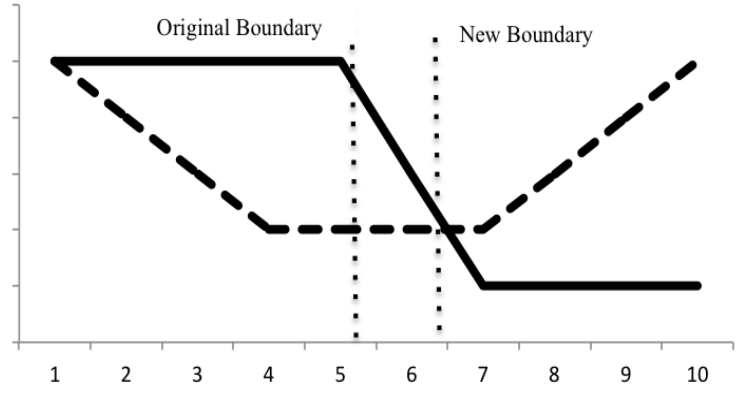


Figure 3.4 depicts the interaction between auditory and visual lexical frequency in producing the McGurk effect. In the main text this interaction is tested in a logit analysis with 1 = non-audio responses and 0 = audio-responses, here we show the effect as 1 minus the average non-audio identifications for each item for ease of reading.

Figure 3.5


Phonetic Identifications & Goodness Ratings
Phonetic Boundary
1 2 3 4 5 6 7 8 9 10 11
Stmulus Item


Predictions: Lexical Context Influences Phonetic Catagorization
Original Boundary
New Boundary
1 2 3 4 5 6 7 8 9 10

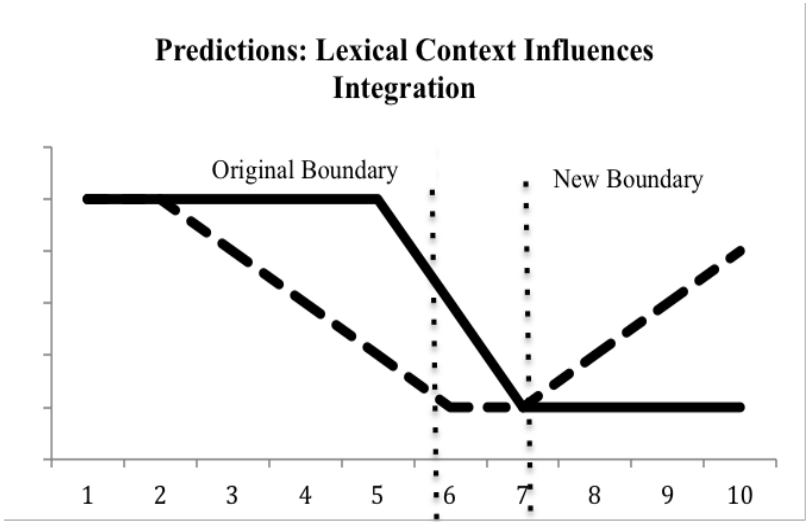

Predictions: Lexical Context Influences Integration
Original Boundary
New Boundary
1 2 3 4 5 6 7 8 9 10

Figure 3.5 illustrates the relationship between goodness ratings and the phonetic boundary (5a) and how effects during categorization (5b) and integration (5c) should influence the goodness scores and phonetic boundary.

Figure 3.6

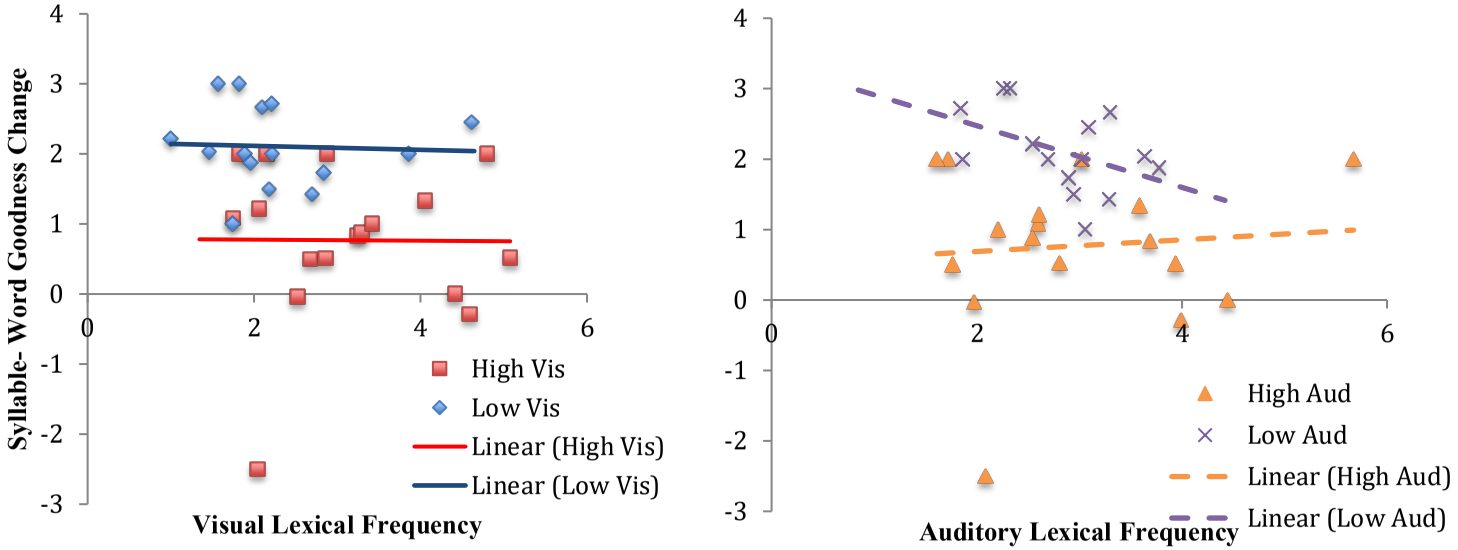


Figure 3.6 shows the interaction between visual word (Panel A) and auditory word (Panel B) lexical frequency and the change in goodness rating between syllable and word blocks. Data points are item averages for items that were not identified as consistent with the auditory or visual (or fusion word) initial consonants during either the syllable and word blocks; these were the points which showed the interaction noted in the discussion. The trend lines in each panel show the items associated with high and low rated syllables. Using the high (syllable average > 2.5) vs. low distinction provides an approximation of the interaction centered on items with a 3-4 out of 5 goodness rating.

Discussion of Dissertation Findings

The results from this dissertation provide several findings relevant to our understanding of multisensory and lexical processes in speech perception. While each chapter in this dissertation addresses a different set of empirical findings, they share a common focus on localizing the effects of lexical information during speech processing relative to multisensory integration. In the following sections we will summarize the key findings and the conclusions drawn from each chapter, before discussing how all the results converge on a set of conclusions.

Chapter 1

Ostrand et al., (2016) reported semantic priming associated with McGurk primes was consistent with the auditory, but not the putatively perceived, component of that prime. From this finding, Ostrand et al., (2016) proposed that semantic processing of the auditory component of the McGurk prime began before multisensory integration completed. However, our investigation presented in Chapter 1 supports a very different conclusion; that semantic priming does in fact reflect the perceived, and visually influenced, component of the prime.

This conclusion is supported by three chief findings. First, the analysis of the lexical decision task showed that semantic priming was consistent with the visual component of the McGurk primes. As the primes used in this experiment were expected to produce more consistent McGurk effects than those used by Ostrand et al., (2016), the results of this experiment can be viewed as being more representative of the relationship

between perception and lexical processing. Second, an analysis of covariance on the McGurk items revealed an interaction between the size of the semantic priming effect and the strength of the McGurk effect for each item. Finally, the interaction found in the analysis of covariance was driven by a positive correlation between the rate of visual word identifications and the size of the priming effect to the McGurk visual word.

A final piece of evidence provided by Chapter 1 comes from a separate study that measured the identification rates for the McGurk words actually used by Ostrand et al., (2016) in their study. This experiment revealed that the McGurk items used by Ostrand et al., (2016) produced lower McGurk rates than did the McGurk items used in our experiment. Using the identification rates from these Ostrand et al., (2016) stimuli and the actual item-level semantic priming data from Ostrand et al., (2016), we found a positive correlation between McGurk rates and semantic priming. This correlation was found to be comparable to the one found for the McGurk stimuli used in our own study. It therefore appears that the results of Ostrand et al's (2016) lexical decision task, like the lexical decision task of our own experiment, reflect semantic priming consistent with the perceived and visually-influenced word.

Chapter 2

Samuel and Lieblich (2014) discuss a series of studies that show how selective adaptation can be driven by lexical context illusions, such as the phonemic restoration effect (Warren, 1970; Samuel, 1997) and the Ganong effect (Ganong, 1980; Samuel, 2001; Samuel & Frost, 2015). They also discuss how selective adaptation is *not* influenced by multisensory illusions such as the McGurk effect (Roberts & Summerfield,

1981; Saldana & Rosenblum, 1994). From these results, these authors argue that speech supports two parallel processes. The first process is linguistic; this is the process that is assumed to drive selective adaptation, as well as any cognitive processes associated with the meaning of a speech segment. The second process is perceptual, and is assumed to determine only the phenomenological experience of a speech stimulus, but not interact at any point with the linguistic process.

In Chapter 2, we began by arguing that a more parsimonious explanation for the dissociation between selective adaptation from lexical and multisensory illusions is that only the multisensory illusions have implemented stimuli that include *clear and conflicting* auditory and visual speech information. The results presented in Chapter 2 show that when this crossmodal conflict is removed, multisensory context can in fact support selective adaptation. Moreover, across experiments, our results seem to suggest that in some ways, this multisensory selective adaptation effect is more robust than the selective adaptation effects associated with lexically-supported phonemic restoration. These results suggest that if selective adaptation does reflect processing of a fundamental and inherently linguistic process, then that process must be sensitive to multisensory information.

Chapter 3

Chapter 3 focuses on work done by Brancazio (2004) who studied lexical effects on the identification of McGurk words. The principal finding from this work is that McGurk effects are more common when they produce word, as opposed to nonword, identifications. In contrast to the studies addressed in chapters 1 and 2, the conclusions

offered by Brancazio (2004) are relatively agnostic; he discusses multiple interpretations of his data but admits that his results are inconclusive. However, Brancazio (2004) also provides an important framework for understanding when during speech processing lexical information might influence speech identification; during multisensory integration or during the phonetic categorization that follows integration.

The work presented in Chapter 3 expands on Brancazio's (2004) work in three ways. First, Experiment 1 of this chapter demonstrates that, to an extent, the identification of McGurk words is predictable from their lexical frequency. This experiment found that the lexical frequency of both of the auditory and visual sensory inputs interact to predict the resulting identification. Second, Experiment 2 expands on the framework offered by Brancazio (2004), adding a stage for lexical influences on perception that takes place prior to multisensory integration, the production of to-be-perceived speech. The critical finding from Experiment 2 is that even when isolated from their word contexts, the identification of McGurk syllables are predicable from the lexical frequency of the words from which they were extracted. Finally, Experiment 3 of Chapter 3 compares the qualitative evaluations of McGurk identifications between syllables presented in isolation and in word contexts. The results of this comparison suggest that lexical processing influences the phonetic categorization, but not the integration, stage of multisensory speech perception.

Conclusions

Despite following different methodologies, the results of all three chapters presented in this dissertation converge in showing that lexical processing *is* sensitive to

the results of multisensory integration. In Chapter 1, semantic priming was consistent with either the auditory or visual channel of a McGurk stimulus, depending on how that stimulus was perceived. In Chapter 2, selective adaptation, which is generally accepted to reflect low level perceptual processing (e.g. see Samuel, 1986), was found to be consistent with multisensory information, even when lexical context failed to support adaptation. Finally, Chapter 3 shows that while lexical effects on speech production can affect multisensory integration, lexical *processing* during perception likely occurs after integration is complete. Taken together these results are consistent with theories of speech perception that assume multisensory integration occurs early and is independent of top-down processing.

The results of this dissertation also converge in supporting a ubiquity of multisensory integration. Across all of our experiments there was no instance in which lexical context, but not multisensory context, could support speech processing. Indeed, in Chapter 2 we found that multisensory context is more reliable than lexical context in some circumstances. This conclusion is consistent with theories that assume the multisensory nature of speech perception is supported by lawful informational correspondences in the multisensory stimuli.

References

Brancazio, L. (2004). Lexical influences in audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(3), 445–463. http://doi.org/10.1037/0096-1523.30.3.445

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology. Human Perception and Performance*, *6*(1), 110–125. http://doi.org/10.1037/0096-1523.6.1.110

Ostrand, R., Blumstein, S. E., Ferreira, V. S., & Morgan, J. L. (2016). What you see isn't always what you get: Auditory word signals trump consciously perceived words in lexical access. *Cognition*, *151*, 96–107. http://doi.org/10.1016/j.cognition.2016.02.019

Roberts, M., & Summerfield, Q. (1981). Audiovisual presentation demonstrates that selective adaptation in speech perception is purely auditory. *Perception & Psychophysics*, *30*(4), 309–314. http://doi.org/10.3758/BF03206144

Saldaña, H. M., & Rosenblum, L. D. (1994). Selective adaptation in speech perception using a compelling audiovisual adaptor. *The Journal of the Acoustical Society of America*, *95*(6), 3658–3661. http://doi.org/10.1121/1.409935

Samuel, A. G. (2001). Knowing a word affects the fundamental perception of the sounds within it. *Psychological Science*, *12*(4), 348–51. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11476105

Samuel, A. G., & Frost, R. (2015). Lexical support for phonetic perception during nonnative spoken word recognition. *Psychonomic Bulletin & Review*, (1970). http://doi.org/10.3758/s13423-015-0847-y

Samuel, A. G., & Lieblich, J. (2014). Visual speech acts differently than lexical context in supporting speech perception. *Journal of Experimental Psychology. Human Perception and Performance*, *40*(4), 1479–90. http://doi.org/10.1037/a0036656

Samuel, A. G. (1986). Red herring detectors and speech perception: In defense of selective adaptation. *Cognitive Psychology*, *18*(4), 452–99. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3769426

Samuel, A. G. (1997). Lexical activation produces potent phonemic percepts. *Cognitive Psychology*, *32*(2), 97–127. http://doi.org/10.1006/cogp.1997.0646

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*(3917), 392–393. http://doi.org/10.1126/science.167.3917.392